Seleksjon av flygere og flygeledere

Innen flytrafikk kan menneskelige svikt føre til katastrofer. Riktig utvelgelse av flygere og flygeledere er avgjørende. Psykologiske tester benyttes for å kartlegge personlige egenskaper, kognitive og psykomotoriske evner.

TEKST
Monica Martinussen

PUBLISERT 1. april 2005

ABSTRACT:
Selection of pilots and air traffic controllers
The purpose of this article was to provide a summary of selection methods of pilots and air traffic controllers. Early tests were initially compiled to select pilots for WW1. Psychological tests designed for ATCO selection were introduced only after WW2. The sustained interest in this area is probably caused by concerns for safety as well as the need for reducing costs associated with training and the operation of aircrafts. Different techniques have been developed to document the predictive validity, including local validation studies and meta-analyses. Meta-analysis suggests that work sample tests and ability tests are in fact valid predictors of pilot and air traffic controller performance. Personality tests however, seem to have a lower and, in some cases, doubtful predictive validity. Methodological problems related to research are discussed.

EMNER flygere seleksjon validitet

Cockpit F-16. Foto: Nils Lio/Forsvarets mediesenter.

Et norsk C130 Herculesfly under et transportoppdrag til Bagram i Afghanistan. Foto: Lars Aamodt/CCT/Forsvarets mediesenter.

Innledning

For 100 år siden kastet brødrene Wilbur og Orville Wright kron og mynt om hvem som skulle få foreta den første flyturen med deres selvbygde flymaskin (Hunter & Burke, 1995). Siden den gang har utvelgelsen av flygere endret seg fra myntkast til å ta i bruk moderne teknologi og databaserte tester. Det er knapt noen psykologisk test som ikke på ett eller annet tidspunkt har vært utprøvd på denne yrkesgruppen. Dette skyldes nok både økonomiske hensyn og tanken om at en nøye utvelgelse av flygere ville kunne bedre sikkerheten og redusere antallet uhell. I tillegg er det store utgifter forbundet

med utdanningen av flygere, innkjøp og drift av flymaskiner. Selv med små forbedringer i frafallsraten under opplæring og færre uhell, vil betydelige beløp kunne være spart. Dette forutsetter selvsagt at de metodene man anvender har prediktiv validitet, og at man dermed treffer flere riktig besluttinger enn ved å kaste kron og mynt. De første valideringsstudiene for å undersøke dette ble gjennomført allerede under første verdenskrig, altså ikke mange år etter at brødrene Wright foretok sin første flytur (Dockeray & Isaacs, 1921).

Det er vanskelig å tenke seg vårt samfunn uten fly og muligheten til å reise lange avstander på kort tid. Et annet aspekt ved luftfart er selvsagt den strategiske betydningen militære fly og flygere har i en konfliktsituasjon. En annen yrkesgruppe som er avgjørende for sikker og effektiv avvikling av flytrafikken, er flygeledere. Utvelgelsen av søkere til flygelederutdanning skjer i de fleste vestlige land ved hjelp av psykologiske tester, men forskningen er langt mindre omfattende enn for flygere (Edgar, 2002). De første psykologiske testene ble tatt i bruk tidlig på 60-tallet og bestod av papir- og blyant-tester (Hätting, 1991). I dag er databaserte tester tatt i bruk, og utvelgelsen er i mange land like omfattende som for flygere.

Både flygere og flygeledere arbeider i et høyteknologisk miljø som stadig endrer seg etter som ny teknologi innføres. I tillegg øker flytrafikken mange steder. Konsekvensene av feil kan for begge yrkesgruppene være katastrofale, i tillegg til å ha store økonomiske konsekvenser. Det er derfor rimelig å vektlegge seleksjon. Formålet med denne artikkelen er å gi en oversikt over hvilken forskning som er gjort, og peke på noen utfordringer som gjenstår når luftfarten går inn i sitt andre århundre.

Flygerseleksjon i et historisk perspektiv

Testutviklingen og de aller fleste valideringsstudiene har blitt utført i militær regi og er basert på søkere til flygerutdanning i forsvaret (Martinussen, 1996). Under første verdenskrig var det et stort behov for å utdanne mange flygere og for å redusere det omfattende frafallet under trening. En rekke tester ble utviklet, og mange av disse var enkle konstruksjoner som skulle simulere oppgaver eller påvirkninger mennesket utsettes for i en flymaskin. Ett av de tidligste testbatteriene utviklet i USA (Henmon, 1919) inneholdt tester som skulle måle emosjonell stabilitet, reaksjonstid, generelle evner og oppfattelse av likevekt.

I Europa ble tilsvarende tester utvik-let i flere land, og i Danmark arbeidet Alfred Lehman (Thermøhlen, 1986) med å utvikle tester til dette formålet i sitt laboratorium. Han foreslo tester som målte emosjonell stabilitet, bedømmelse av romlige forhold, oppmerksomhet, reaksjonstid for lyd, og oppfattelse av likevekt. Testen som målte emosjonell stabilitet besto av enkelte psykofysiologiske målinger samtidig som at man uventet avfyrte et skudd bak ryggen på kandidaten. Lehman mente at testen var uegnet til seleksjon fordi det ikke var mulig å skille de som var virkelig kaldblodige fra de som reagerte på stress indusert i testsituasjonen (Thermøhlen, 1986). Det var mange likheter mellom testene som ble brukt i de ulike landene i denne første fasen av testutvikling.

Papir- og blyant-tester var i bruk sammen med apparater som etterlignet aspekter ved en flymaskin.

Det ble tidlig framhevet at også personlige egenskaper var viktige for å kunne bli en god flyger, i tillegg til kognitive og psykomotoriske evner. For å kartlegge hvilke personlighetstrekk som var aktuelle, ble både observasjon av flygere og deltakende observasjon benyttet. Dockeray konkluderte med, etter selv å ha gjennomgått flygerutdanning, at «Quiet methodological men were among the best flyers, that is, the power and quick adjustment to a new situation and good judgment» (Dockeray & Isaacs, 1921).

Etter at første verdenskrig var avsluttet, var det liten forskningsaktivitet på flygerseleksjon i de fleste land (Hilton & Dolgin, 1991). Ett unntak var Tyskland der en rekke tester ble utviklet, og ved starten på 2. verdenskrig hadde de et testbatteri som bestod av hele 29 tester som målte generell intelligens, perseptuelle evner, koordineringevne, karakter og lederegenskaper (Fitts, 1946). I løpet av krigen ble dette testbatteriet erstattet av et mindre omfattende system med færre tester og mer vekt på referanser og intervjudata (Fitts, 1946). I England, USA og Canada var utviklingen motsatt, og ved starten av krigen var få tester i bruk, mens mot slutten av krigen var en rekke tester utviklet og i bruk. I Norge ble de første testene tatt i bruk av Luftforsvaret i 1946 (Riis, 1986). Siden den gang er det norske testbatteriet utviklet og validert flere ganger (se f.eks. Martinussen & Torjussen, 1998; Torjussen & Hansen, 1999).

Etter 2. verdenskrig var det igjen en lengre periode med redusert forsk-ningsinnsats i forhold til å utvikle og validere nye tester. I USA ble det startet et omfattende program for å finne egnede personlighetsmål for flygerseleksjon. Forskningsprogrammet ble ledet av Saul Sells (1955, 1956), og hele 26 personlighetsmål ble vurdert. Sells og kollegaer brukte mer langsiktige kriterier på flygerprestasjoner enn det som hadde vært vanlig. De konkluderte med at personlighetsmål var bedre prediktorer på lengre sikt sammenlignet med evnetester der den prediktive validiteten avtok over tid.

En rekke velkjente personlighetsinventorier har også blitt utprøvd, som MMPI (Melton, 1954), Eysenck Personality Inventory (Bartram & Dale, 1982; Jessup & Jessup, 1971), Rorschach (Moser, 1981) og Cattell 16PF (Bartram, 1995a). Resultatene viste kun lave til moderate korrelasjoner med kriteriet. I Sverige utviklet Ulf Kragh (1960) en projektiv test kalt «The Defence Mechanism Test» (DMT). Formålet var å velge ut søkere til høyrisikoyrker, for eksempel flygere. Selve testmaterialet består av TAT-lignende bilder som presenteres ved hjelp av et tachistoskop. Eksponeringstiden er meget kort, men øker for hver gang bildet presenteres. Personen skal tegne og fortelle hva han eller hun ser, og avvik mellom det faktiske bildet og det som personen rapporterer blir da tolket som ulike forsvarsmekanismer. Dette er en svært forenklet framstilling av en meget komplisert og omfattende skåringsprosedyre (se f.eks. Torjussen & Værnes, 1991). Testen ble møtt med betydelig optimisme da den ble lansert, og den ble prøvd ut på militære flygere i flere land som England, Nederland og Australia i tillegg til i Skandinavia (Martinussen & Torjussen, 1993). Imidlertid har det vært vanskelig å dokumentere testens prediktive validitet for flygere ut over i de skandinaviske landene,

og testen anvendes i dag kun i liten grad til dette formålet (English & Rodgers, 1992; Martinussen & Torjussen, 1993; Turnbull, 1992).

Med innføringen av datamaskiner i testingen har en rekke andre personlighetsrelaterte begreper blitt testet ut. Dette har vært mål på risikotaking, selvsikkerhet, feltavhengighet og ulike holdningsmål (se Hunter & Burke, 1995 for en oversikt). I de fleste tilfellene har dette kun resultert i meget små korrelasjoner med kriteriet, og ingen økning i den prediktive validiteten utover det som evnetestene predikerte. Nyere studier har imidlertid gitt mer positive resultater for trekkbaserte personlighetsmål for flygerseleksjon (Bartram & Baxter, 1996; Hörmann & Mascke, 1996). Personlighet vektlegges i dag i varierende grad under utvelgelsen. Enkelte land som USA og England bruker ikke personlighetstester i sin seleksjon (Carretta & Ree, 2003), men en vurdering av personlige egenskaper og motivasjon kan selvsagt komme inn som et moment under intervjuet med psykolog eller offiser.

I tillegg til evne-og personlighetstester har også andre typer av prediktorer vært prøvd ut. Både biografiske data, kunnskap om yrket og tidligere flygererfaring har vært benyttet. I dag anvender mange land en trinnvis seleksjonsprosess der både enkle og mer komplekse kognitive tester anvendes sammen med mål på motivasjon. Fordelen med en slik trinnvis prosess er at man slipper å teste alle kandidatene med alle testene, og kostnadene reduseres dermed.

De første databaserte testene så dagens lys på 70- og 80-tallet (Bartram, 1995b; Hunter & Burke, 1987; Kantor & Carretta, 1988). Etter hvert som datateknologien ble både billigere og bedre, ble papir- og blyant-testene erstattet helt eller delvis med databaserte tester i de fleste vestlige land (Burke et al., 1995).

«Når tester skal anvendes til seleksjon, er det avgjørende at man kan dokumentere at testene faktisk predikerer framtidige prestasjoner»

Utvelgelse av flygeledere

De fleste valideringsstudiene som er gjennomført i forhold til seleksjon av flygeledere, er utført av de amerikanske luftfartsmyndighetene (FAA) (Collins, Boone, & Deventer, 1980; Sells, Daily, & Pickerley, 1984). De første testbatteriene som ble tatt i bruk på 60-tallet, inneholdt papir- og blyant-tester som målte resonering (verbal og numerisk), samt perseptuell hurtighet og spatiale evner (Hätting, 1991). Testresultatene ble sammen med utdanning, alder og erfaring brukt i seleksjonen. På 70-tallet startet FAA utviklingen av en simulatorbasert test som skulle måle kandidatenes ferdigheter i å anvende ulike regler i et simulert luftrom. Testen kom senere i en papir- og blyant-utgave, og fikk navnet «Multiplex Controller Aptitude Test». Den ble brukt sammen med mål på resonnering og yrkeserfaring i utvelgelsen fra starten på 80-tallet (Carretta

& Siem, 1999). Utviklingen av et databasert testbatteri startet på 90-tallet. Dette målte spatial resonnering, korttidshukommelse, oppfattelse av bevegelse, mønster-gjenkjenning og oppmerksomhet (Broach & Manning, 1997).

I Europa gjennomførte Eurocontrol på slutten av 70-tallet (Hätting, 1991) en gjennomgang av medlemslandene sine seleksjonsprosedyrer, og de fleste anvendte tester som målte spatial persepsjon, verbale evner, resonnering og hukommelse. Få land brukte tester for å kartlegge interesse eller motivasjon for yrket. Et unntak var Tyskland der German Aerospace Center i tillegg til et omfattende testbatteri, også brukte et mål på personlighetstrekk og en simulatorbasert test for å måle samarbeidsevne (Eißfeldt, 1991, 1998).

På 80-tallet startet utviklingen av databaserte tester i både Tyskland (Hätting, 1991) og England (Burke, 1992). I 2003 lanserte EUROCONTROL et felles databasert testbatteri for utvelgelsen av flygeledere som medlemslandene kan anvende. I dag er altså papir- og blyant-testene erstattet helt eller delvis av databaserte tester i mange land. Både basale kognitive funksjoner kartlegges i tillegg til at man anvender datateknologien til å simulere deler av arbeidsoppgavene som en flygeleder har. I Sverige har man lagt ned et betydelig arbeide i å utvikle et standardisert intervju, kalt situasjonsintervju, for å kartlegge både enkelte evner og sosiale holdninger (Brehmer, 2003). I Norge anvendes ulike databaserte tester, simulerte arbeidsoppgaver og intervju i en trinnvis prosess (Martinussen, 2003). I tillegg stilles det i alle land medisinske krav, og formelle krav til alder og utdanning.

Hvordan evaluere testenes prediktive validitet?

Når tester skal anvendes til seleksjon, er det først og fremst avgjørende at man kan dokumentere at testene faktisk predikerer framtidige prestasjoner. For å dokumentere dette kan man enten utføre en lokal valideringsstudie, eller vise til meta-analyser der det foreligger evidens for at en gitt type tester har prediktiv validitet i forhold til et yrke eller en arbeidsoppgave. Lokale valideringsstudier skjer vanligvis ved at man korrelerer testresultatene med ett eller annet mål på arbeidsprestasjoner, for eksempel prestasjoner i en simulator eller vurderinger gjort av en instruktør. I slike studier er det viktig at kriteriene på arbeidsprestasjoner både er relevante for organisasjonen, har god begrepsvaliditet og reliabilitet. I mange tilfeller kan det være vanskelig å utføre lokale valideringsstudier fordi man ikke ansetter mange nok personer innenfor en tidsperiode til at utvalget blir stort nok.

Meta-analyse

Et alternativ til å utføre en lokal valideringsstudie er å kombinere tidligere undersøkelser i en meta-analyse. Meta-analyse er en fellesbetegnelse på at man anvender statistiske teknikker til å integrere resultater fra mange undersøkelser. For at man skal kunne slå sammen resultater fra flere undersøkelser, må disse finnes på en felles målestokk. De to mest brukte indeksene er produkt-moment-korrelasjon (r) og effekt-størrelse $(ES = (M_1\text{-}M_2)/SD)$. Begge indeksene angir hvor stor effekt man har

funnet, og dermed hvor godt det er mulig å predikere en variabel ut fra kjennskap til den andre variabelen.

Begrepet meta-analyse ble lansert på slutten av 70-tallet av Glass (1976) og kollegaer som arbeidet med psykoterapiforskning. Samtidig publiserte Schmidt og Hunter (1977) sine første arbeider om validitetsgeneralisering, dvs. i hvilken grad testers prediktive validitet kan generaliseres over ulike situasjoner. Er det slik at intelligenstester alltid kan brukes til å predikere jobb- og skoleprestasjoner, eller er det slik at en intelligenstest må undersøkes hver gang den skal brukes til et slikt formål?

Hunter og Schmidt (1990) utviklet sin egen utgave av meta-analyse, der hovedformålet var å beregne den gjennomsnittlige prediktive validiteten samt å undersøke variasjon mellom undersøkelser i prediktiv validitet. I de fleste formene for meta-analyse legges det stor vekt på å beregne en gjennomsnittlig effekt, mens de skiller seg når det gjelder hvordan variasjon mellom studier skal undersøkes. Enkelte foretrekker signifikanstesting, dvs. at man tester hypotesen om at variasjonen mellom studier er statistisk signifikant forskjellig fra null. Dersom dette er tilfelle, antar man at det er faktiske forskjeller mellom studiene som må studeres nærmere. Forkastes ikke nullhypotesen antar man, i hvert fall inntil videre, at det ikke er noen sann variasjon mellom studiene. Den beregnede gjennomsnittlige effekten er dermed et godt mål på den faktiske prediktive validiteten til en type test. Hunter og Schmidt (1990) foreslo i stedet for signifikanstesting at man anslår den sanne variansen mellom studier ved å beregne differansen mellom observert varians mellom studier og varians som skyldes tilfeldige feil («sampling error»).

Statistiske feilkilder i valideringsstudier

Hunter og Schmidt (1990) har beskrevet en rekke faktorer eller forhold som kan påvirke størrelsen på den observerte korrelasjonen (eller effektstørrelsen). Tre slike statistiske feilkilder er mang-lende reliabilitet, redusert spredning i variablene («restriction of range»), og bruk av et todelt kriterium i stedet for et kontinuerlig. Jo dårligere reliabiliteten til variablene er, desto lavere blir den observerte korrelasjonen. Dette er det mulig å korrigere for dersom man kjenner til variablenes reliabilitet. Dersom man ønsker å korrigere for reliabiliteten til begge variablene, er formelen (Hunter & Schmidt, 1990):

$$r_{cor} = \frac{r_{obs}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

Bokstavene rcor er den korrigerte korrelasjonen, robs er den observerte korrelasjonen, rxx og ryy er reliabiliteten til hver av variablene. Den korrigerte korrelasjonen gir altså korrelasjonen som ville ha blitt observert dersom variablene hadde blitt målt med perfekt reliabilitet. I en del tilfeller er det bare aktuelt å korrigere for reliabilitet med hensyn til den ene av variablene. For eksempel der man skal bruke tester til å predikere

senere prestasjoner. I dette tilfellet er det bare aktuelt å korrigere for reliabiliteten til prestasjonsmålet. Dette fordi man ønsker å finne ut nytten av testene med de feil og mangler som de måtte ha. Korreksjonen blir da:

$$r_{cor} = \frac{r_{obs}}{\sqrt{r_{yy}}}$$

Den andre faktoren som også påvirker størrelsen på korrelasjonen er redusert spredning i skårene på den ene eller begge variablene som følge av seleksjon («restriction of range»). Et eksempel er at man bare studerer sammenhengen mellom testresultater og senere prestasjoner hos de som er valgt ut på bakgrunn av sine testprestasjoner. Kanskje er bare den beste halvdelen av gruppen valgt ut, og det er disse man har muligheten til å innhente kriteriedata på. Den beregnede korrelasjonen blir da langt lavere for denne gruppen enn dersom vi hadde undersøkt hele gruppen. Effekten på den observerte korrelasjonen kan være tildels dramatisk, gitt at seleksjonen er streng nok. Et eksempel hentet fra 2. verdenskrig, viste at den prediktive validiteten til samlet skåre («Pilot stanine») basert på ulike evnetester var .64 for militære flygere. Dersom man hadde anvendt informasjonen fra testene og selektert den 13 % beste andelen av kandidatene, ville den prediktive validiteten blitt .18 (Thorndike, 1949). Dette gir et bilde av den til dels dramatiske effekten av å beregne den prediktive validiteten basert på en strengt selektert gruppe. Korreksjonen skjer på bakgrunn av informasjon om spredningen i hele gruppen, eller andelen som er selektert av hele søkergruppen (Hunter & Schmidt, 1990). I de tilfeller der seleksjon er basert på flere tester, eller på tester anvendt i kombinasjon med annen informasjon, blir situasjonen mer komplisert, og mer avanserte modeller for korreksjon bør anvendes (Johnson & Ree, 1994; Lawley, 1943).

En tredje feilkilde er at man anvender et dikotomt kriterium i stedet for et kontinuerlig mål på prestasjoner. Dette gjelder ofte i valideringsstudier av både flygere og flygerledere der bestått/ikke bestått utdanning er et vanlig mål på prestasjon. Dette fører i likhet med dårlig reliabilitet og redusert spredning i kriterieskårene til en lavere korrelasjon mellom test og kriterium, enn dersom man hadde anvendt et kontinuerlig kriterium. Også denne feilkilden er det mulig å korrigere for hvis man vet forholdet mellom antallet bestått/ikke bestått (Hunter & Schmidt, 1990).

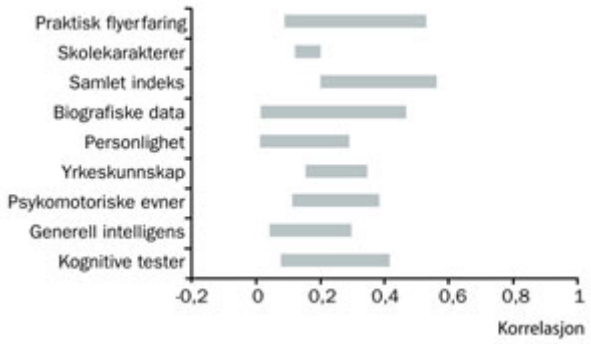
Dersom studier varierer i forhold til reliabilitet eller hvor streng seleksjonen er, vil dette bidra til at man observerer forskjeller i prediktiv validitet mellom studiene. Hunter og Schmidt (1990) foreslår derfor at korrelasjonene bør korrigeres for dette før de slåes sammen i en meta-analyse. Dessverre er det få studier som korrigerer for disse feilkildene, eller rapporterer de opplysningene som skal til, for å foreta slike korreksjoner.

Når kan testvaliditeten generaliseres?

Hvordan kan man så avgjøre om testvaliditeten til en type prediktor kan generaliseres over ulike situasjoner? Hunter og Schmidt (1990) har foreslått en tommelfingerregel som sier at dersom minst 75 % av den observerte variansen mellom korrelasjoner kan tilskrives statistiske feilkilder og utvalgsfeil («sampling error»), er det rimelig å anta at den øvrige variansen skyldes feilkilder som man ikke har klart å korrigere for. Dermed antar man den sanne variansen mellom studier er svært liten eller null. Den andre situasjonen oppstår når man har en viss sann populasjonsvarians, og man kan da anslå et intervall som den prediktive validiteten befinner seg innenfor. Dette intervallet beregnes med utgangspunkt i den korrigerte gjennomsnittlige korrelasjonen og det estimerte populasjonsstandardavviket (Whitener,1990). Dersom dette intervallet er stort og i tillegg inneholder null, innebærer det at den faktiske variasjonen mellom studier er betydelig, og at testen i noen tilfeller ikke har prediktiv validitet. Andre ganger kan det være slik at intervallet er av en viss størrelse, men at det ikke inkluderer null. Dette betyr at det er en viss variasjon i den prediktive validiteten, men at den alltid er positiv.

Hvor godt virker testene?

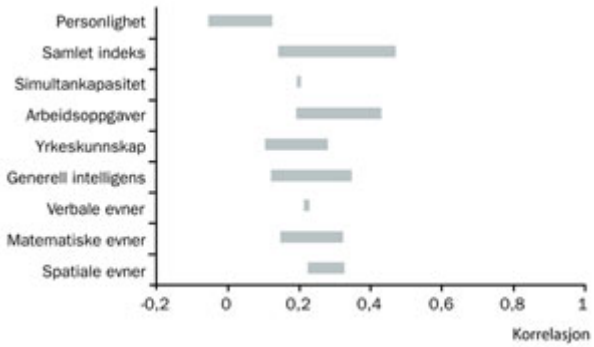
Det er utført langt færre valideringsstudier når det gjelder tester brukt til å velge ut flygeledere enn flygere. I 2000 ble det utført en meta-analyse av tilgjengelige studier, og man fant da til sammen 25 artikler/rapporter som dokumenterte valideringsdata for 35 ulike utvalg (Martinussen, Jenssen, & Joner, 2000). Studiene var publisert mellom 1952 og 1999, og de fleste var basert på søkere og studenter som tok flygelederutdanning (92 %). De fleste undersøkelsene var gjennomført i USA (77 %), og kriteriene som ble benyttet bestod stort sett av vurderinger innhentet under utdanning (bestått/ikke bestått, instruktørvurderinger, simulator). Resultatene fra disse 25 artiklene ble kombinert i en meta-analyse der den gjennomsnittlige prediktive validiteten ble beregnet og den faktiske variansen til testene ble estimert. De samlede utvalgene varierte mellom 224 og 11 255 personer for de ulike kategorier av prediktorer. Praktisk talt ingen av studiene rapporterte informasjon som gjorde det mulig å korrigere for manglende reliabilitet i kriteriet og redusert spredning i teskårene («restriction of range»). De gjennomsnittlige korrelasjonene er derfor et underestimat av den sanne prediktive validiteten. Korrelasjonene ble korrigert for bruk av et dikotomt kriterium. De ulike testene og prediktorene ble gruppert, og en oppsummering av disse resultatene er presentert i Figur 1. Det skraverte feltet indikerer området for sann varians der de 10 % mest ekstreme verdiene på hver side av fordelingen er kuttet ut (kredibilitetsintervall) (Whitener, 1990).



Figur 1. Prediktiv validitet for ulike tester anvendt for utvelgelse av flygere basert på resultater fra meta-analyse (Martinussen, 1996).

For alle, unntatt to av prediktorene (verbale evner og simultan-kapasitet), er det en viss sann varians. Nedre grense for intervallet ligger over null for alle kategorier unntatt personlighetstester. Dette indikerer at selv om det er en viss variasjon i den prediktive validiteten, er den positiv over utvalg og situasjoner med unntak av personlighetstester.

Når det gjelder flygere, er det gjennomført langt flere valideringsstudier. Det er publisert flere litteraturgjennomganger (se f.eks. Carretta & Ree, 2003; Hunter, 1989) og to meta-analyser som summerer opp forskningen på området (Hunter & Burke, 1994; Martinussen, 1996). På tross av et noe forskjellig datagrunnlag og enkelte prosedyreforskjeller i måten meta-analysene var gjennomført på, viste de to studiene svært like hovedfunn. En oversikt over de estimerte intervallene (kredibilitetsintervall) er presentert i Figur 2. Figuren er basert på 50 studier der resultater for 66 ulike utvalg er inkludert (Martinussen, 1996). De samlede utvalgene for ulike testgrupper varierte mellom 3736 og 17900 personer. Som for flygeledere, var det ikke mulig å korrigere de observerte korrelasjonene for annet enn effekten av å anvende et dikotomt kriterium.



Figur 2. Prediktiv validitet for ulike tester anvendt for utvelgelse av flygeledere basert på resultater fra meta-analyse (Martinussen, Jenssen, & Joner, 2000).

Diskusjon

Resultatene av meta-analysen (Martinussen, 1996; Martinussen, Jenssen, & Joner, 2000) viste for både flygere og flygerledere at man langt på vei har lyktes i å utvikle kognitive tester som kan anvendes i seleksjonen. Når det gjelder personlighetsmål er resultatene langt mindre imponerende. Dette kan skyldes flere faktorer, og betyr ikke at personlighet ikke spiller noen rolle i yrkesutøvelsen. Tvert i mot viser jobbanalyser for både flygere og flygerledere at personlige egenskaper er viktige for å gjøre en god jobb. For flygeledere fremheves samarbeidsevne, gode kommunikasjonsferdigheter, og evne til å takle stress (Eißfeldt & Heintz, 2002). For militære flygere har egenskaper som mestringsmotivasjon, evne til å treffe beslutninger og handle raskt samt emosjonell stabilitet blitt framhevet som særlige viktige (Carretta, Rodgers, & Hansen, 1996). Hva kan så forklaringen være på den manglende prediktive validiteten? En mulighet er at de personlighetstestene som har vært prøvd ut, ikke har vært egnet til seleksjon. For eksempel ved at man har prøvd ut kliniske måleinstrumenter som opprinnelig var laget for å diagnostisere problemer eller patologi. En annen mulighet er at det på selvrapporteringstester er opplagt hva som er et sosialt ønskelig svar, og dermed er det enkelt for søkerne å framstille seg selv i et gunstig lys. En meta-analyse av måleinstrumenter basert på femfaktormodellen viste imidlertid at selv om søkerne i noen grad framstiller seg selv i et gunstig lys, så reduserte dette ikke testenes prediktive validitet nevneverdig (Ones, Viswesvaran, & Reiss, 1996).

«Vi har lyktes i å utvikle kognitive tester som kan anvendes i seleksjonen, men når det gjelder personlighetsmål er resultatene langt mindre imponerende»

Generelt har innføringen av femfaktormodellen for personlighet introdusert en viss optimisme når det gjelder bruk av personlighetsmål til seleksjon (se f.eks. Sandal, 1999). Meta-analyser har vist at trekk som «Samvittighetsfull» og «Emosjonell stabilitet» predikerer jobbprestasjoner i en rekke yrker (Barrick & Mount, 1991; Salgado, 1997; Schmidt & Hunter, 1998). Det er likevel mulig at bruk av hoved-dimensjonene i femfaktormodellen ikke vil gi mye i forhold til yrkesgrupper der man også må regne med en viss selv-seleksjon av søkerne. Muligens vil enkelte av underfaktorene være bedre i forhold til å skulle predikere framtidige arbeidsprestasjoner.

Et annet moment er at de kriteriene som anvendes, ofte er innhentet under opplæringen. Da er det rimelig at enkelte kognitive ferdigheter er mer betydningsfulle enn personlige egenskaper. Dette er i tråd med funnene til Sells (1955, 1956), som indikerte at personlighetstester var bedre prediktorer på litt sikt. Det er imidlertid få valideringsstudier som anvender faktiske jobbprestasjoner. Dette kan skyldes at det kan være praktisk vanskelig å innhente kriteriedata når personene arbeider ulike steder. Et annet moment er at personer under utdanning vil kunne vurderes under mer eller

mindre like forhold, fordi de avlegger de samme prøvene og gjør de samme øvelsene. Når utdanningen er avsluttet, blir arbeidet mer forskjellig. Det blir dermed vanskeligere å finne et kriterium som er likt for alle. I tillegg kan det være at arbeidstakere motsetter seg å skulle bli vurdert under arbeidet selv om formålet er å undersøke testenes validitet og ikke finne feil hos den enkelte.

Det er publisert svært få studier om utvelgelse av flygere til sivile flyselskap. Dette skyldes flere faktorer. Dels har mange flyselskap basert seg på å ansette flygere som allerede har sin utdanning, ofte fra forsvaret, og dermed vil en del testing være unødvendig. Et annet moment er et behov for å beskytte egne seleksjonsprosedyrer fra både søkere og fra konkurrenter i markedet. Et siste moment er at mange flyselskaper ikke ansetter et stort nok antall flygere om gangen, eller har ressurser til å utføre valideringsstudier.

Databaserte tester

Det har vært nedlagt et betydelig arbeide i å utvikle og validere tester for utvelgelse av både flygere og flygeledere. Trenden har gått fra papir- og blyant-tester og mekaniske psykomotoriske tester til databaserte testbatteri. Bruken av databaserte tester har gjort det mulig å teste mer komplekse og sammensatte evner og ferdigheter enn tidligere. Det er mulig å måle reaksjonstid og oppmerksomhet både alene og som et aspekt av en mer omfattende oppgave. Det er også mulig å simulere deler av de framtidige arbeidsoppgaver som flygerleder, og det er mulig å presentere informasjon både på skjermen og gjennom hodetelefoner. Innføringen av databaserte tester har medført forenklinger i form av at administrasjon og skåring av testene blir enklere. Samtidig må datamaskiner og programvare oppdateres slik at også databaserte tester krever vedlikehold.

Et problem ved slike komplekse dynamiske tester er at testen kan utvik-le seg forskjellig for de ulike søkerne. Valg og prioriteringer man gjør på et tidlig tidspunkt kan få konsekvenser senere for både arbeidsbelastningen og kompleksiteten på oppgaven. I tillegg kan det være at søkerne anvender ulike evner og strategier for å løse oppgavene. Noen kan for eksempel prioritere hurtighet framfor sikkerhet og nøyaktighet. Dette gjør skåringen av slike tester mer komplisert enn ved enklere tester der man kan summere opp antall riktige oppgaver. Et annet moment ved slike dynamiske tester er at de gjerne krever en lengre instruksjon og innføring før selve testingen kan ta til. Dette gjør dem tidkrevende, og ofte anvendes disse på et senere tidspunkt i seleksjonen der søkergruppen er allerede er testet med enklere tester og de svakeste søkerne er skilt ut.

Bruk av PC gjør det også mulig å anvende en mer tilpasset testing, dvs. at vanskegraden på oppgavene bestemmes av hvordan vedkommende utfører de første oppgavene i testen. Denne type tester er basert på såkalt «item respons»-teori (se f.eks. Embretson & Reise, 2000) der formålet er å anslå personens evnenivå. Ofte er disse testene kostbare i utviklingsfasen, og de fleste testene som anvendes i dag er basert på klassisk test-teori.

Med Internett er det nå blitt mulig å teste søkere som kan befinne seg hvor som helst, og man sparer da reiseutgifter. Problemet med dette er hvordan man skal sikre seg at det

faktisk er søkeren som gjennomfører selve testen. Et annet moment er at dette kan gjøre det vanskeligere å holde informasjon om testinnholdet hemmelig. Dermed blir det vanskeligere å hindre at søkere gjør seg kjent med testene og øver på dem før de søker opptak. Likevel er det trolig at en slik pre-testing vil bli tatt i bruk av flere. Dermed kan søkere undersøke selv om dette er noe som passer for dem. På det neste trinnet i seleksjonen vil søkerne måtte innkalles til en videre testing der man kan forsikre seg om at det er riktig person som blir testet, og man kan sjekke at personen faktisk pres-terer så godt som resultatene innhentet via Internett antyder.

En bekymring som har vært reist i forhold til bruk av databaserte tester, er om de er ekvivalente til papir- og blyant-utgaven av den samme testen. Måler man det samme begrepet, og vil de som testes rangeres på samme måte uavhengig av måten testen administreres på? Flere studier har undersøkt sammenhengen mellom papir -og blyant-tester og databaserte utgaver av de samme testene. En meta-analyse av evnetester viste generelt høyt grad av samsvar mellom ulike versjoner av samme test, men for tester der tiden er den avgjørende faktoren, var forskjellen noe større (Mead & Drasgow, 1993). En grunn til at man finner forskjeller avhengig av administrasjonsmodus kan være at enkelte har mer erfaring med bruk av datamaskiner og spill. Det er ikke dermed gitt at testene får en dårligere prediktiv validitet selv om det er forskjeller mellom testene avhengig av administrasjonsmodus. En undersøkelse blant flygerledere viste at dataerfaring i seg selv predikerte jobbkriterier etter at man hadde kontrollert for testprestasjoner (Young, Broach, & Farmer, 1997). En studie av databaserte tester for norske flygere i forsvaret viste en sammenheng mellom dataerfaring og utførelse på de databaserte testene som målte psykomotoriske ferdigheter (Eide, 1999). For å motvirke dette problemet er det derfor vanlig å legge inn en del øvelser slik at søkerne skal bli godt kjent med tastatur og annet utstyr som anvendes før selve testingen tar til. En annen mulighet er å gjøre visse designmessige endringer i forhold til tastatur og stikke slik at de blir mer ulike det som vanligvis brukes i forhold til spill. Dermed blir fordelen mindre for de med mye spill- og dataerfaring.

Luftfartens andre århundre: Hva vil framtiden bringe?

Det er avgjørende at valg av tester til seleksjon er basert på empiri om hva som faktisk predikerer framtidige jobbprestasjoner, med andre ord en evidensbasert praksis. Oppsummeringen av forskningen viste at det er behov for mer langsiktige studier av testenes prediktive validitet og bruk av kriterier ut over skoleprestasjoner. I tillegg er det viktig at man korrigerer for de aktuelle statistiske feilkildene, slik at man får et mest mulig riktig bilde av den faktiske nytten til testene. Når det gjelder ulike kognitive tester og arbeidsoppgaver («work sample tests»), har man for både flygere og flygerledere utviklet gode tester. Med utviklingen av nyere testteori («item respons»-teori) og bedre datateknologi, vil vi helt sikkert få se både nye testtyper og mer effektive måter å teste på. Antakelig vil også bruk av testing via Internett bli brukt i en tidlig fase i seleksjonen. I forhold til valg av personlighetstester er det klart rom for forbedringer, og en del av forskningen vil trolig bygge på femfaktormodellen for personlighet.

Monica Martinussen

Institutt for psykologi

Universitetet i Tromsø, 9037 Tromsø

Tlf 77 64 43 48

E-post moncam@psyk.uit.no

Teksten sto på trykk første gang i Tidsskrift for Norsk psykologforening, Vol 42, nummer 4, 2005, side 291-300

TEKST

Monica Martinussen, Avdeling for militærpsykologi og lederutvikling, Forsvarets Høgskole

Referanser

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1-26.

Bartram, D. (1995a). The predictive validity of the EPI and 16PF for military flying training. Journal of Occupational and Organizational Psychology, 68, 219-236.

Bartram, D. (1995b). Validation of the MICROPAT battery. International Journal of Selection and Assessment, 3, 83-94.

Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airwaiys selection program. The International Journal of Aviation Psychology, 6, 149-169.

Bartram, D., & Dale, H. C. A. (1982). The Eysenck Personality Inventory as selection test for military pilots. Journal of Occupational Psychology, 55, 287-296.

Brehmer, B. (2003). Predictive validation of the MRU Battery. Proceedings of the Second EUROCONTROL selection seminar. (HRS/MSP-002-REP-07). Brussels: EUROCONTROL.

Broach, D., & Manning, C. A. (1997). Review of air traffic controller selection: An international perspective (DOT/FAA/AM-97/15). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Burke, E. F. (1992). Validity of the Controller Index. Unpublished report. London: Ministry of Defence.

Burke, E., Kokorian, A., Lescreve, F., Martin, C. J., Van Raay, P., & Weber, W. (1995). Computer-based assessment: A NATO survey. International Journal of Selection and Assessment, 3, 75-83.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. I P. S. Tsang & M. A. Vidulich (Eds.), Principles and practice of aviation psychology (ss. 357-396). New Jersey: Lawrence Erlbaum Associates.

Carretta, T., Rodgers, M. N., & Hansen, I. (1996). The identification of ability requirements and selection instruments for fighter pilot training. Technical report 2 from Euro-Nato Aircrew Human Factor Working Group.

Carretta, T. R., & Siem, F. M. (1999). Determinants of enlisted air traffic controller success. Aviation, Space, and Environmental Medicine, 70, 910-918.

Collins, W. E., Boone, J. O., & VanDeventer, A. D. (1980). The selection of air traffic control specialists: History and review of contributions by the Civil Aeromedical Institute, 1960-1980. Aviation, Space, and Environmental Medicine, 52, 217-240.

Dockeray, F. C., & Isaacs, S. (1921). Psychological research in aviation in Italy, France, England, and the Americal Expeditonary Forces. Journal of Comparative Psychology, 1, 115–148.

Edgar, E. (2002). Cognitive predictors in ATCO selection: Current and future perspectives. I H. Eißfeldt, M. C. Heil & D. Broach (Eds.), Staffing the ATM system (ss. 73–83). Aldershot, England: Ashgate.

Eide, T. A. (1999). Seleksjon av flygere. Innføring av databasert testbatteri ved Luftforsvarets seleksjonssenter: Har dataerfaring betydning for testskåren? Hovedoppgave levert ved Psykologisk institutt, Universitetet i Oslo.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah; New Jersey: Lawrence Erlbaum Associates.

English, A., & Rodgers, M. (1992). Déjà vu? Cultural influences on aviator selection. Military Psychology, 4, 35–47.

Eißfeldt, H. (1991). DLR selection of air traffic control applicants. I E. Farmer (Ed.), Human resource management in aviation (ss. 37–49). Aldershot: Avebury Technical.

Eißfeldt, H. (1998). The selection of air traffic controllers. I K. M. Goethers (Ed.), Aviation psychology: A science and a profession (ss. 73–80). Aldershot, England: Ashgate.

Eißfeldt, H., & Heintz, A. (2002). Ability requirements for DFS controllers – Current and future. I H. Eißfeldt, M. C. Heil & D. Broach (Eds.), Staffing the ATM system (ss. 13–24). Aldershot, England: Ashgate.

Fitts, P. M. (1946). German applied psychology during World War 2. American Psychologist, 1, 151–161.

Glass, G. V. (1976). Primary, secondary, and meta-analysis research. Educational Researcher, 5, 3–8.

Hätting, H. J. (1991). Selection of air traffic control cadets. I R. A. Galand & A. D. Mangelsdorff (Eds.), Handbook of military psychology (ss. 115–148). Chichester: John Wiley & Sons.

Henmon, V. A. C. (1919). Air service tests of aptitude for flying. The Journal of Applied Psychology, 2, 103–109.

Hilton, T. F., & Dolgin, D. L. (1991). Pilot selection in the military of the free world. I R. Gal & A. D. Mangelsdorff (Eds.), Handbook of military psychology (ss. 81–101). New York: John Wiley.

Hunter, D. R. (1989). Aviator selection. I M. F. Wiskoff, & G. M. Rampton (Eds.), Military personnel measurement (ss. 29–167). New York: Praeger.

Hunter, D. R., & Burke, E. F. (1995). Handbook of pilot selection. Aldershot, England: Avebury Aviation.

Hunter, D. R., & Burke, E. F. (1987). Computer-based selection testing in the Royal Air Force. Behaviour Research Metods, Instruments, & Computers, 19, 243–245.

Hunter, D. R., & Burke, E. F. (1994). Predicing aircraft pilot training: A meta-analysis of published research. The International Journal of Aviation Psychology, 4, 297–313.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research. Newbury Park: Sage Publications.

Hörmann, H., & Maschke, P. (1996). On the relation between personality and job performance of airline pilots. The International Journal of Aviation Psychology, 6, 171–178.

Jessup, G., & Jessup, H. (1971). Validity of the Eysenck Personality Inventory in pilot selection. Journal of Occupational Psychology, 45, 111–123.

Johnson, J. T., & Ree, M. J. (1943). Rangej: A Pascal program to compute the multivariate correction for range restriction. Educational and Psychological Measurement, 54, 693–695.

Kantor, J. E., & Carretta, T. R. (1988). Aircrew selection systems. Aviation, Space, and Environmental Medicine, 59, A32–A38.

Kragh, U. (1960). The Defense Mechanism Test: A new method for diagnosis and personnel selection. Journal of Applied Psychology, 44, 303-309.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae: Proceedings of the Royal Society of Ediburgh, 62 (section A, part 1), 28-30.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. The International Journal of Aviation Psychology, 6, 1-20.

Martinussen, M. Jenssen, M., & Joner, A. (2000). Selection of air traffic controllers: Some preliminary findings from a meta-analysis of validation studies. Proceedings from the 24th EAAP (European Association for Aviation Psychology) conference.

Martinussen, M. (2003). Controller selection: Recent validation studies in Norway. EUROCONTROL selection seminar. HRS/MSP-002-REP-07. Brussels: EUROCONTROL.

Martinussen, M., & Torjussen T. (1993). Does DMT (Defense Mechanism Test) predict pilot performance only in Scandinavia? I R. S. Jensen & D. Neumeister (Eds.), Proceedings of the Seventh International Symposium on Aviation Psychology (ss. 398-403). Columbus: Ohio State University.

Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. The International Journal of Aviation Psychology, 8, 33-45.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerbased and paper-and pencil cognitive ability tests. A meta-analysis. Psychological Bulletin, 119, 449-458.

Melton, R. S. (1954). Studies in the evaluation of the pesonality characteristics of successful naval aviators. Journal of Aviation Medicine, 25, 600-604.

Moser, U. (1981). Eine Methode zure Bestimmung Wiederstandsfähigkeit gegenüber der Konfliktreaktiverung unter Verwendung des Rorschachtests, dargestellt am Problem der Pilotenselektion. Schweizerische Zeitschrift für Psychologie, 40, 279-313.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679

Riis, E. (1986). Militærpsykologien i Norge. Tidsskrift for Norsk Psykologforening, 23 (Suppl. 1), 21-37.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. Journal of Applied Psychology, 82, 30-43.

Sandal, G. M. (1999). Personlighetstester som rekrutteringsmetode i næringslivet. Tidsskrift for Norsk Psykologforening, 36, 764-771.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124, 262-274.

Sells, S. B. (1955). Development of a personality test battery for psychiatric screening of flying personnel. Journal of Aviation Medicine, 26, 35-45.

Sells, S. B. (1956). Further developments on adaptibility screening for flying personnel. Aviation Medicine, 27, 440-451.

Sells, S. B., Dailey, J. T., & Pickrel, E. W. (1984). Selection of air traffic controllers. (FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Termøhlen, J. (1986). Flyvepsykologiens udvikling. I I. K. Moustgaard & A. F. Petersen (Red.), Udviklingslinier i dansk psykologi: Fra Alfred Lehmann til i dag (ss. 169-181). København: Gyldendal.

Thorndike, R. L. (1949). Personnel selection. New York: Wiley.

Torjussen, T. M., & Hansen, I. (1999). Forsvaret: Best i test? Bruk av psykologiske tester i forsvaret, med spesiell vekt på flygerseleksjon. Tidsskrift for Norsk Psykologforening, 36, 772-779.

Torjussen, T. M., & Værnes, R. (1991). The use of the Defence Mechanism test (DMT) in Norway for selection and stress reseach. I M. Olff, G. Godaert & H. Ursin (Eds.), Quantification of human defence mechanisms. Berlin: Springer Verlag.

Turnbull, G. J. (1992). A review of military pilot selection. Journal of Aviation, Space, and Environmental Medicine, 63, 825–830.

Young, W. C., Broach, D., & Farmer, W. L. (1997). The effects of computer-based experience on air-traffic controller specialists, air traffic scenario test scores (DOT/FAA/Am-97/4). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. Journal of Applied Psychology, 75, 315–321.