NTNU

Kunnskap for en bedre verden

Utvelgelse av kandidater til Sjøforsvarets grunnleggende befalskurs

En undersøkelse av den prediktive validiteten til seleksjonssystemet

Tore Norrøne

Hovedoppgave i psykologi

Trondheim, februar 2016

Norges tekniske-naturvitenskaplige universitet
Fakultet for samfunnsvitenskap og teknologiledelse
Psykologisk institutt

Forord

Først og fremst ønsker jeg å takke psykologene på Forsvarets høgskole. Tom, Joachim, Ole Christian og Jan – dere har vært en kilde til oppmuntring og støtte fra første øyeblikk! Jeg kunne ikke bedt om bedre kolleger.

En velfortjent takk rettes også til mine veiledere, Eva og Monica. Deres faglige interesse og øye for detaljer var inspirerende og motiverende i en prosess hvor inspirasjon og motivasjon var sårt trengt.

Til slutt vil jeg takke deg, Thea. For at du hele tiden minner meg på at det er viktigere ting i livet enn jobb.

Innhold

Sammendrag

Formålet med denne studien var å undersøke den predikative validiteten til seleksjonssystemet ved Sjøforsvarets grunnleggende befalskurs. Dette ble gjort ved å undersøke sammenhengen mellom de ulike prediktorene og prestasjoner ved endt utdanning, både som elev og militær leder for en gruppe på 8-10 personer. Det var også ønskelig å se hvor mye av variansen i prestasjonene opptaket forklarte og hvorvidt hver seleksjonstest bidro med unik forklart varians (incremental validity). Prediktorer var GMA (generell mental evne), gjennomsnittspoeng fra videregående skole, intervju (skoleprognose og lederprognose) og vurderingssenter. Korrelasjonsanalyser viste at alle prediktorene var signifikant korrelert enten med prestasjonene som elev, prestasjonene som militær leder eller begge. De sterkeste sammenhengene med prestasjonene som elev var gjennomsnittspoeng fra videregående skole (r = .39), skoleprognosen fra intervjuet (r = .33) og GMA-testene (r = .31). De sterkeste sammenhengene med prestasjonene som militær leder var vurderingssenteret (r = .41) og lederprognosen fra intervjuet (r = .40), mens GMA var ikke signifikant korrelert med dette kriteriet. Regresjonsanalysene viste også at prediktorene samlet forklarte 24 % av variansen i prestasjonene som elev og 28 % av variansen i prestasjonene som militær leder.

Introduksjon

Seleksjon basert på psykologiske tester i det norske Forsvaret har tradisjoner helt tilbake til andre verdenskrig (Eid, Lescreve, & Larsson, 2012). Regjeringen befant seg denne perioden i eksil i London og ble dermed sterkt påvirket av britene og amerikanernes tanker og erfaringer med seleksjon av personell. Det ble opprettet en kommisjon som ga føringer om mer fokus på intelligenstesting når man vurderer nye rekrutter til tjeneste i Forsvaret. Etter andre verdenskrig ble det første psykologiske testsenteret opprettet under ledelse av Forsvarets første Sjefspsykolog. Norge opererte med tvungen førstegangstjeneste for alle menn, noe som var med på å tydeliggjøre behovet for grundig seleksjon og klassifisering av rekruttene. Til sammen 30 000 mann skulle hvert år bli tildelt en rolle hvor de kunne bidra til å styrke landets forsvar under den kalde krigen. Mange psykologiske tester er blitt utviklet og revidert siden denne første perioden i Forsvarets psykologitjeneste, og noen få, slik som sesjonstestene som inngår i målet generell mental evne (GMA), har tålt tidens tann og brukes relativt uendret i Forsvaret fortsatt (Torjussen & Hansen, 1999). På 1970-tallet ble fokuset endret til militær ledelse og seleksjon av offiserer. Dette førte til utviklingen av øvelser som skulle gjøre det mulig å måle og utvikle personlighet, interpersonlige ferdigheter, lederskap og selvinnsikt hos militære ledere. Disse øvelsene, modifisert og rettet mot seleksjon, er bakgrunnen for dagens ordning, et såkalt vurderingssenter (assessment center), som er inkludert i denne studien.

Seleksjon av personell er en svært viktig del av alle organisasjoner, spesielt hvis arbeidsoppgavene er vanskelige å gjennomføre, eller kritiske at blir gjort korrekt. Rett person på rett plass er med på å øke produksjonen i en organisasjon betraktelig, og bør derfor være regnet som et svært gunstig satsningsområde sett fra et økonomisk perspektiv (Hull,1928; Hunter & Hunter, 1984; Judiesch & Schmidt, 2000; Terpstra & Rozell, 1993). I tillegg til

dette kan en uprofesjonell seleksjonsmetode eller en ikke-eksisterende seleksjonsmetode gi søkeren en negativ oppfatning av organisasjonen, noe som senere kan skade organisasjonen i form av et dårlig rykte (Andersen, 2004). Slike hendelser kan også ende med at kandidaten formelt klager på organisasjonen. I en organisasjon som Forsvaret er det heller ikke utenkelig at man havner i situasjoner hvor feil avgjørelse kan være direkte knyttet til tap av eget eller andres liv, noe som gjør at seleksjonen er desto mer viktig. Det er heller ikke kun organisasjonen som drar nytte av god seleksjon. Det kan tenkes å være svært ubehagelig å bli satt i en stilling hvor man ikke innehar de nødvendige ferdighetene, evnene eller egenskapene som trengs for å mestre den, spesielt når det er situasjoner som potensielt er farlige for liv og helse. Man kan med andre ord argumentere for at en organisasjon som Forsvaret har et ansvar overfor sine kandidater for å gjennomføre en grundig og profesjonell seleksjon. Med dette menes seleksjon med et sterkt empirisk fundament, ettersom seleksjon basert på eksperters intuisjon har vist seg å være lite treffsikkert (Dakin & Armstrong, 1989).

Militære organisasjoner har tradisjonelt vært ansett som ledende innenfor seleksjon av personell (Driskell & Olmstead, 1989). De viktigste grunnene til dette er at de militære organisasjonene er delt inn i avskilte enheter med relativt lite intern forflytning, et høyt antall personer vurderes hvert år og det er vanlig å gå rett ut i en opplæringsfase som gir en nær kobling mellom vurdering og trening (Rumsey, 2012). Denne strukturen, i tillegg til å underbygge behovet for klassifisering og seleksjon av personell, er med på å gi en god arena for validering av metodene som blir benyttet.

Kvalitetssikring av seleksjonsmetoder – validitet og feilkilder

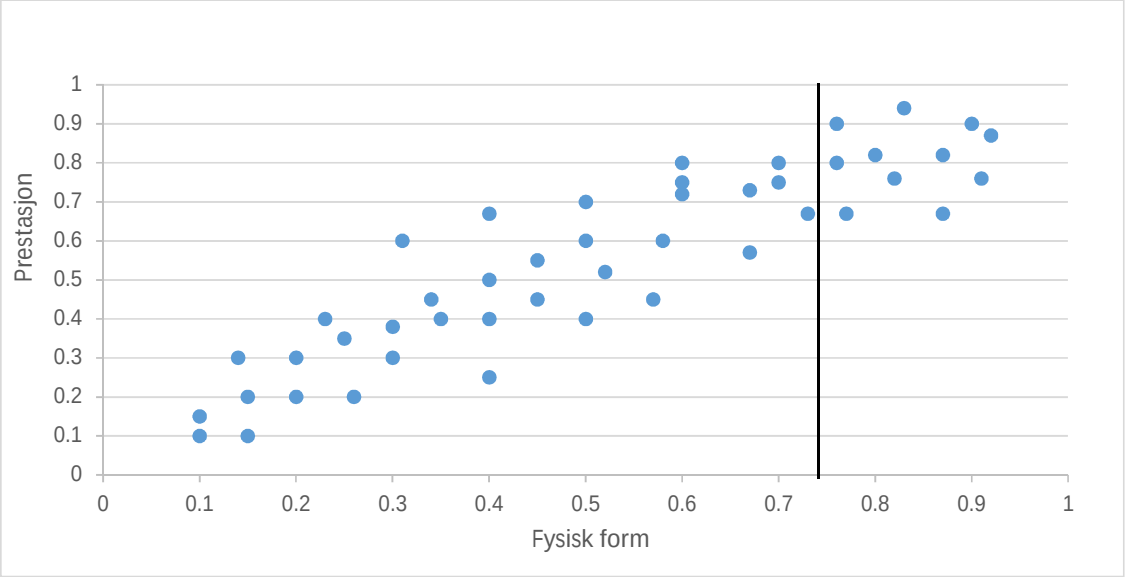
En vanlig måte å kvalitetssikre en seleksjonsprosedyre på er ved å gjennomføre en lokal valideringsstudie. En valideringsstudie innenfor seleksjon fokuserer på hvorvidt, og i hvilken grad, testen eller andre prediktorer korrelerer med jobbprestasjonen i den aktuelle stillingen (kriterievaliditet). Det er en pragmatisk form for validitet, fordi den konkret

besvarer spørsmålet «hvor godt fungerer seleksjonen?», uten å si noe om hvorfor den fungerer. Denne studien vil også i stor grad benytte seg av meta-analyser for å argumentere for hvor godt ulike metoder predikerer jobbprestasjoner. En meta-analyse er en kombinasjon resultater fra flere primærstudier, hvor det benyttes statistiske teknikker til å summere opp resultatene (Martinussen, 2005). Fokuset vil ligge på studier som bruker produkt-moment korrelasjon (*r*), da resultater fra valideringsstudier ofte rapporteres i form av dette. Gjennom meta-analyse teknikker er det mulig å beregne en gjennomsnittlig korrelasjon basert på en rekke studier og blant annet studere i hvilken grad validiteten lar seg generalisere over ulike utvalg og situasjoner. Et viktig i aspekt i slike valideringsstudier er valget av kriterium som bør kunne måles både reliabelt og med god begrepsvaliditet. I tillegg er det selvsagt ønskelig at man anvender et kriterium som oppleves som relevant for organisasjonen. Vanlige kriterier i slike studier er vurderinger gjort av overordnede eller resultater fra opplæring/utdanning.

Tilleggsvaliditet (*incremental validity*) sier noe om seleksjonsmetoden bidrar til at forklart varians i kriteriet øker utover det som allerede er forklart ved de metodene som allerede er i bruk. Dette er viktig å vurdere fordi en test kan ha høy kriterievaliditet, men fortsatt være bortkastet tid og ressurser hvis den ikke øker den forklarte variansen av kriteriet. Det er også aktuelt å snakke om *korrigerte validitetskoeffisienter*. Dette begrepet ble utformet fordi man tenkte at kriterievaliditeten som ble beregnet i seleksjonsstudier ofte var kunstig lav på grunn av ulike metodiske feilkilder i studiene. Dette var feilkilder som begrenset variasjon i testskårene, som følge av seleksjon, og manglende reliabilitet i kriteriet (Schmidt & Hunter, 2003).

Hvis Marinejegerkommandoen i Forsvaret bestemmer seg for å vurdere hvorvidt det hjelper å være i god fysisk form for å mestre deres utdanning, og tester dette ved å korrelere fysisk form med prestasjoner i tjenesten, vil de sannsynligvis ikke finne så sterke korrelasjoner. Dette til tross for at disse variablene åpenbart bør henge sammen. Grunnen til

dette er at alle soldatene i et slik avdeling er nøye utvalgt og i svært god fysisk form, så spredningen i fysisk form er begrenset. Det at redusert spredning påvirker styrken på korrelasjonen er lett å vise visuelt. Figur 1 viser en fiktiv korrelasjon mellom de to variablene og det er lett å se den stigende linjen (positiv korrelasjon) i datasettet hvis man ser hele bildet. Hvis man derimot kun fokuserer på datapunktene til høyre for den svarte vertikale linjen, blir det lineære bildet raskt mer sirkulært (lav eller ingen korrelasjon). Den svarte linjen i grafen representerer opptakskravet til Marinejegerkommandoen slik at alle søkerne med lavere skåre er valgt bort.



Figur 1: Betydningen av begrenset variasjon i seleksjon

I seleksjon vil man alltid møte problemet med begrenset variasjon, fordi man ikke har mulighet til å se hvordan de man selekterte bort ville prestert hvis de fikk muligheten. Det eneste man har mulighet til å vurdere er treffsikkerheten på rangeringen av de man faktisk selekterte. Et unntak på dette er et eksempel rapportert av Thorndike (1949). Under andre verdenskrig sendte USA en uselektert gruppe inn i pilottrening på grunn av stor mangel på piloter og kort forventet levetid. Dette medførte at 77 % av kandidatene strøk før treningen var over. Den beregnede korrelasjonen mellom totalskåren basert på alle testene og

bestått/ikke-bestått trening var .64. Hadde man anvendt testene på vanlig måte og kun selektert de beste ville korrelasjonen vært .18 (Thorndike, 1949). For andre studier er det mulig å foreta en statistisk korreksjon av korrelasjonen som gir den teoretiske korrelasjonen man ville ha hatt uten problemet med redusert spredning. Korreksjonen forutsetter at man kjenner til spredningen i hele søkergruppen eller andelen som selekteres (Schmidt & Hunter, 2003). Er seleksjonen en funksjon av flere variabler og en kombinasjon av både målte og ikke målte variabler (for eksempel inntrykk fra et intervju) er det mer komplisert å korrigere for denne feilkilden.

Ved de aller fleste valideringsstudier slik som denne ønsker man å teste resultatene opp mot kriteriet «jobbprestasjon». Dette gir mening fordi organisasjoner er interessert i å finne «den mest kompetente arbeider» som bidrar positivt til sin nye arbeidsplass. Spørsmålet blir da hvordan skal man bedømme jobbprestasjon? Den mest brukte måten å vurdere jobbprestasjon på er vurderinger fra overordnet (Lent, Aurbach, & Levin, 1971). Det er åpenbare problemer med denne måten å vurdere jobbprestasjoner på, blant annet at forskjellige ledere eller overordnede ikke alltid ser samme person på samme måte, eller at ikke alle ledere har like god kjennskap til sine ansatte og hvordan de presterer. En meta-analyse fant at inter-rater reliabiliteten til slike vurderinger var $r = .57$ (Viswesvaran, Ones, & Schmidt, 1996). Dette er en noe lav reliabilitet og indikerer at selv om en seleksjonstest predikterer jobbprestasjon, så vil ikke nødvendigvis valideringsstudier klare å avdekke dette fordi lav reliabilitet i kriteriet vil redusere den beregnede korrelasjon mellom prediktor og kriterium. Løsningen på dette er, på samme måte som ved begrenset variasjon, en statistisk korreksjon av korrelasjonen, en metode som har er blitt kritisert for å føre til overestimering av kriterievaliditeten (LeBreton, Scherer, & James, 2014). En slik korrigering forutsetter at man kjenner reliabiliteten til kriteriet. En annen måte å måle jobbprestasjon på er i form av

karakterer under utdanning. For denne typen jobbprestasjonsmål er inter-rater reliabiliteten beregnet til $r = .80$ (Hunter & Hunter, 1984), noe som er langt mer akseptabelt.

Rettferdige metoder og indirekte diskriminering

Et viktig aspekt ved seleksjon i tillegg dokumentasjon av prediktiv validitet er at metodene som anvendes er rettferdige og ikke diskriminerer noen grupper (Evers et al., 2013). Når man snakker om indirekte diskriminering i seleksjon snakker man som regel om seleksjonsmetoder som resulterer i færre ansettelser av et bestemt kjønn eller etnisk minoritet enn det de representert i søkergruppen, men det kan også innebære andre karakteristika som ikke bør bli vektlagt i en profesjonell utvelgelse (for eksempel søkerens seksuelle orientering). Dette skiller seg fra direkte diskriminering ved at det ikke er intensjonelt, men heller en bieffekt av testen eller seleksjonssystemet. Et eksempel på dette er når man selekterer basert på plettfri vandel, slik som Forsvaret gjør. Hvis det da viser seg at noen grupper oftere har en anmerkning på rullebladet sitt enn majoriteten, fører dette til indirekte diskriminering. I slike situasjoner er det viktig at organisasjonen har et bevisst forhold til dette, slik at de kan begrunne seleksjonsmetoden på en saklig og god måte (International Test Commission, 2001). Dette vil som regel bety at man viser til metodens prediktive validitet og koblingen mellom metoden og jobbanalysen. Hvis man ikke kan vise til noe evidens for hvorfor akkurat denne testen skal brukes i akkurat denne seleksjonsprosessen, risikerer man å tape klagesaker eller søksmål som omhandler nettopp dette. For en organisasjon som Forsvaret, som jobber eksplisitt og målrettet for å øke andel kvinner og etniske minoriteter i organisasjonen, kan et slikt utfall være svært uheldig for omdømmet.

Problemet med å sikre både høy validitet og lav indirekte diskriminering er at disse variablene har en tendens til å korrelere positivt, noe som medfører at de mest valide testene også er de med høyest indirekte diskriminering (Ployart & Holtz, 2008). For Forsvaret sin del er det spesielt prediktorer som gjennomsnittskarakterer (Roth & Bobko, 2000) og GMA

(Roth, Bevier, Bobko, Switzer, & Tyler, 2001b) som i andre studier har vist indirekte diskriminering, men dette er også vist seg å være et problem for ustrukturerte intervju (Huffcutt, Conway, Roth, & Stone, 2001) og vurderingssenter (Dean, Roth, & Bobko, 2008).

Hvordan velge prediktorer?

En jobbanalyse er et naturlig grunnlag å ha før seleksjonsprosessen planlegges, og det er et viktig utgangspunkt for å sikre treffsikker seleksjon (Morgeson & Campion, 2000). Jobbanalysen går ut på å systematisk analysere bredden og dybden i arbeidsoppgavene som tilhører stillingen som analyseres. Ettersom alle andre avgjørelser knyttet til seleksjon baserer seg på hvordan stillingen er definert, vil en dårlig eller ikke-eksisterende jobbanalyse kunne føre til følgefeil som svekker troverdigheten til seleksjonsprosessen. En god jobbanalyse vil også kunne forbedre rekruttering, trening og utvikling, bruk av ansatte og det vil fungere som et juridisk dokument ved eventuelle klagesaker eller søksmål (Cook, 2009). Et eksempel på jobbanalyse er å identifisere hvilke oppgaver som innehaveren av stillingen møter ofte, hvilke oppgaver som er spesielt viktige og hvilke oppgaver som er de mest komplekse.

Når arbeidsoppgavene er kartlagt og dimensjonert har man et solid grunnlag for å vurdere hvilke kompetanser som skal til for å prestere godt i denne stillingen. Ved fravær av en jobbanalyse er en slik kompetanseliste ofte basert på en overfladisk ekspertvurdering der resultatet er usikkert. Tre metaanalyser har vist at personlighetstesting (Tett, Jackson & Rothstein, 1991), strukturerte intervjuer (Wiesner & Cronshaw, 1988) og «Situational Judgement Tests» (McDaniel, Hartman, & Grubb, 2007) har høyere validitet når de er valgt ut basert på en jobbanalyse. Det er også viktig at listen av kompetanser ikke blir for lang. Det vil for eksempel være svært vanskelig å vurdere mange kompetanser i løpet av et kort intervju. Dette kan føre til at vurderingen byttes ut med mindre treffsikre spørsmål, som for eksempel «Føles det som at dette er en dyktig person?» eller «Liker jeg denne personen?» (Kahneman & Frederick, 2002).

I Forsvaret brukes kompetansemodellen som er avbildet i Figur 2. Denne er utarbeidet med utgangspunkt i kompetansemodellene til Lai (2004) og Skorstad (2015), og sier noe om hvilke kompetansetyper som er viktigere enn andre i en seleksjonsprosess. Tanken bak disse ulike kompetansetypene er at det er viktigere å selektere basert på de kompetansene som er vanskelig å endre, enn de som er mer fluktuerende. Dette er fordi en intelligent person med de rette personlighetstrekkene vil kunne lære seg de nødvendige kunnskapene og ferdighetene, mens heving av intelligens eller endring av personlighet i beste fall være svært omfattende og tidkrevende, om ikke umulig. Det er eksempler på når ferdigheter og kunnskap av praktiske årsaker bør være grunnleggende i seleksjonsprosessen. Det kan være når det er omfattende å lære opp de nyansatte (for eksempel hvis kravet er at du må være jurist) eller hvis stillingen er så oversiktlig og ferdighets- eller kunnskapsspesifikk at få spesifikke mål er tilstrekkelig (for eksempel bueskytter). Motivasjon på sin side er regnet som relativt lett å endre. I denne modellen er det derfor overlatt mer til utdanningen eller stillingen i samarbeid med den utvalgte å finne tilstrekkelig motivasjon til å yte, nettopp fordi det er lettere å motivere noen i en stilling enn å selektere basert på motivasjon. Dette henger sammen med at det er svært vanskelig å predikere motivasjon. En kandidat som har levd hele livet med drømmen om å bli militært befal er ikke nødvendigvis mer motivert i stillingen/utdanningen enn en kandidat som tilfeldigvis bestemte seg for å søke 30 minutter før søknadsfristen gikk ut. Hvis vurderingen i tillegg skal tas basert på et intervju kan det være vanskelig å skille mellom motivasjon, ekstroversjon eller karisma. Det er også rimelig å anta at handlingen av å søke på en stilling, blandet med gjennomføringen av krevende seleksjonstester, antyder en viss grad av motivasjon.

Figur 2: Forsvarets kompetansemodell

Per dags dato har ikke Forsvaret gjort en jobbanalyse av stillingene som møter elevene etter endt utdanning, men heller en analyse av hvilke kompetanser som kreves for å være en god militær leder. Denne ble gjort på 90-tallet ved at hærens kompetansesenter for lederutvikling (HKLU) spurte befalsskolene, utdanningsavdelingene, inspektoratene, Stabsskolen og kadettene fra Krigsskolen om hvilke kompetanser de legger vekt på i lederutdanning. Dette, sammen med litteraturstudier, dannet grunnlaget for fem kjernekompetanser (Stokke, 2000). De fem kjernekompetansene som definerer en god militær leder er: en som takler usikkerhet, skaper tillit, tar selvstendige og gode beslutninger, viser omsorg og tar initiativ. Kompetansene brukes av alle militære grener blant annet under opptaket (intervjuet og vurderingssenteret), under utdanning (skikkethet som militær leder) og ved fremtidige jobber i Forsvaret (strukturerte referanser).

Seleksjonsmetoder i Forsvaret

Når jobbanalysen er på plass og de essensielle kompetansene er identifisert og prioritert ut fra kompetansemodellen, er det ønskelig å se nøyere på seleksjonsmetodene. Ettersom Sjøforsvarets grunnleggende befalskurs bruker intervju, tidligere prestasjoner, GMA og vurderingssenter i sitt opptak, vil disse metodene bli kort oppsummert her. Her vil den generelle og militærspesifikke kriterievaliditeten være i fokus, for å gi et grundig sammenligningsgrunnlag for valideringen av utdanningen.

Tester på generell mental evne (GMA) som seleksjonsmetode

Forsvarets «alminnelig evnenivå» er regnet som et mål på GMA og er godt dokumentert som reliable og valide tester (Sundet, Barlaug, & Torjussen, 2004). En definisjon eller måte å forstå GMA på er som evnen til å lære (Hunter, 1986). I et mer finkornet perspektiv ser man svært generelle mentale prosesser som resonnering, planlegging, problemløsning, abstrakt tenkning, forståelse for komplekse ideer, rask læring og muligheten til å lære av erfaring (Gottfredson, 1997a). GMA er med andre ord mer enn det å være «skoleflink». Det er en generell og grunnleggende evne til forstå omgivelsene og løse de utfordringene som oppstår i disse omgivelsene. Denne tanken om en generell mental evne eller g-faktor er basert på at prestasjoner på kognitive tester har en tendens til å korrelere høyt med hverandre, og har mye støtte innenfor seleksjonsforskning (Hunter & Hunter, 1984; Judge, Bono, Ilies, & Gerhardt, 2002; Judge, Colbert, & Ilies 2004; Robertson & Smith, 2001; Schmidt & Hunter, 1998). GMA har også møtt på del kritikk, både knyttet til hvorvidt den eksisterer (Sternberg & Wagner, 1993) og hvorvidt den er nyttig (Guilford, 1988). Denne studien har en mer pragmatisk interesse hvor det som er viktig er hvorvidt seleksjonen fungerer eller ikke – altså om testene predikerer prestasjoner i jobb senere. I en slik sammenheng er det kriterievaliditeten som er viktigst å fokusere på.

GMA sin evne til å predikere arbeidsprestasjoner er svært godt dokumentert i ulike studier og meta-analyser (Murphy, 2002; Schmidt, 2002). Chamorro-Premuzic og Furnham (2010) oppsummere noen av de mest siterte og robuste studiene i den enorme mengden studier som er gjort for å vise denne koblingen. Disse viser at den gjennomsnittlige validiteten til GMA ligger på r = .40 - .60 korrigert og r = .25 ukorrigert (Hunter & Hunter, 1984; Judge, Higgins, Thoresen, & Barrick, 1999; Schmidt, 2002; Schmidt & Hunter, 1998, 2004). Dette gjelder også for militære utvalg (Melchers & Annen, 2010; Caretta et al., 2014). Det betyr at GMA forklarer omtrent 25 % av variasjonen i jobbprestasjonene som er målt i disse studiene,

hvis man går ut fra den korrigerte validiteten. Oppsummerende sier Ree og Earles (1992) at tester som måler GMA har den beste prediktive validiteten av alle tester som måler psykologiske variabler. Det er også mye støtte for at GMA bør brukes sammen med tester som kartlegger andre sider ved individet, som for eksempel personlighetstester. Flere andre variabler har vist å ha tilleggsvaliditet, blant annet tester som måler trekket planmessighet (conscientiousness) og integritet (Bobko, Roth, & Rotosky, 1999; Schmidt & Hunter, 1998), mens det ustrukturerte intervjuet, til tross for tidligere tvil, er vist å korrelere høyt med GMA (Roth & Huffcutt, 2013). I sum gjør den høye validiteten og den lave utgiften med å administrere slike tester at GMA-tester er en svært kostnadseffektiv seleksjonsmetode.

Det er ikke slik at GMA og jobbprestasjon korrelerer likt for alle yrker. Validiteten til GMA funnet i en meta-analyse å øke med kompleksiteten i jobben (Salgado, Anderson, Moscoso, Bertua, De Fruyt, & Rolland, 2003). En stor metaanalyse av jobber med forskjellig kompleksitet viste at GMA korrelerte signifikant med jobb -og treningsprestasjoner innenfor alle nivå av kompleksitet, men jo mer kompleks jobben/treningen er jo høyere er korrelasjonen. I disse studiene varierte den korrigerte gjennomsnittskorrelasjonen fra $r = .23$ i de minst komplekse yrkene, til $r = .59$ i de mest komplekse (Hunter, 1980; Hunter & Hunter, 1984).

Forskningen som er gjort på alminnelig evnenivå (Forsvarets GMA) viser et tilsvarende bilde som ble funnet i den internasjonal forskningslitteraturen. Kjenstadbakk (2012) undersøkte 194 befalselever ved Hærens befalsskole mellom 2010 og 2012. Blant disse ble det funnet en ukorrigert korrelasjon på .26 mellom alminnelig evnenivå og hovedkarakter etter endt utdanning. Studier av testene som anvendes for å selektere flygere til forsvaret har vist at generelle evnetester så vel som tester som måler mer spesifikke evner predikerer senere prestasjoner (Martinussen & Torjussen, 2004).

Tidligere prestasjoners evne til å prediktere fremtidige prestasjoner

To vanlige måter å bruke tidligere prestasjoner på i seleksjonssammenheng er i form av referanser og karakterer fra fullførte utdanningsløp. Referanser er en svært utbredt metode som brukes av de aller fleste arbeidsgiver i Europa, og er mest utbredt i Skandinavia (Dany & Torchy, 1994). Til tross for dette er referanser en av de prediktorene med lavest validitet av de som brukes mye i dag (Judge & Higgins, 1998). Meta-analysen til Reilly og Chao (1982) ga en gjennomsnittskorrelasjon mellom referanser og jobbprestasjon på $r = .18$ hvis kriteriet var basert på overordnedes vurderinger. Selv ved korrigering for redusert spredning og kriteriets reliabilitet ble korrelasjonen kun på .26 (Hunter & Hunter, 1984). Interessant nok ble den høyeste korrelasjonen funnet i USA sitt sjøforsvar, hvor den korrigerte korrelasjonen ble målt til $r = .36$ (Jones & Harrison, 1982). Til tross for generelt sett svake funn er det grunner til å tro at referanser kan komme med et viktig bidrag i seleksjon, hvis det gjøres på riktig måte. Det er fordi strukturerte referanser er vist å ikke korrelere med GMA, noe som gir tilleggsvaliditet (Zimmerman, Triana, & Barrick, 2010). I tillegg til dette er referanser et av få seleksjonsmål som sier noe om typisk prestasjon, heller enn optimal prestasjon (Taylor, Pajo, Cheung, & Stringfield, 2004).

Vik (2013) gjorde en valideringsstudie av 183 elever på Hærens krigsskole og fant at tjenesteuttalelsene fra tidligere militære stillinger, som er Forsvarets strukturerte referanse, var den sterkeste enkeltprediktoren av gjennomsnittlig karakter ved endt utdanning ($r = .35$, ukorrigert). Dette til tross for at studien også inneholdt mål på personlighetstrekk («Big Five»), intervju og GMA. Også i denne studien hadde tjenesteuttalelsene ingen statistisk signifikant korrelasjon med GMA eller personlighetstesten noe som er med på å underbygge at referanser kan tilføre noe som ikke er kartlagt ved hjelp av tester. En signifikant korrelasjon ble derimot funnet mellom referansene og intervjuet, noe som virker naturlig siden tjenesteuttalelsen var tilgjengelig da resultatet fra intervjuet ble satt. Analysene til Vik (2013) ble gjort på et utvalg av lavere befal. Det er i senere tid gjort studier som indikerer at

validiteten til tjenesteuttalelsene vil være lavere hvis man gjør en lignende analyse på høyere offiserer (Thomassen, 2014).

Karakterer fra tidligere utdanning er også med på å bidra til informasjon om typisk prestasjon (litt avhengig av type utdanning), fordi det handler om å opprettholde prestasjonen over såpass lang tid at man lettere ser den gjennomsnittlige prestasjonen. For Forsvaret sin del er det poengene fra videregående skole som brukes mest. Internasjonal forskning har funnet at GPA (grade-point average) er en god seleksjonsmetode av flere grunner. Først og fremst har metodens korrigerte validitet vist seg i en meta-analyse å ligge på .33 for jobbprestasjoner (Roth, Bevier, Switzer, & Schippmann, 1996), en gjennomsnittskorrelasjon som viste seg å være høyere ett år etter endt utdanning (r = .45, men falt til .11 etter seks år) og høyere ved opptak til utdanning, militære yrker eller finans. Videre er GPA regnet som et robust mål på longitudinell akademisk prestasjon, en god måte å selektere til grunnleggende opptak (Rynes, Orlitzky, & Bretz, 1997), i tillegg til å være svært praktisk fordi informasjonen er lett tilgjengelig og krever ingen ekstra testing.

Tidligere norske studier som ønsket å utforske hvorvidt poeng fra videregående skole predikterte prestasjon i militære utdanningsløp støtter opp under funnene til Roth et al. (1996). Blant disse er de ukorrigerte korrelasjonene r = .27 opp mot gjennomsnitt vitnemål etter endt krigsskole (Vik, 2013), r = .26 opp mot hovedkarakter etter fullført befalsskole i Hæren (Kjenstadbakk, 2012) og r = .36 opp mot fullført befalsskole i Luftforsvaret (Isaksen, 2014).

Intervju som seleksjonsmetode

I motsetning til GMA og referanser, er intervjuet en svært utbredt metode. Med unntak av Tyrkia, som gjennomfører intervju i 64 % av tilfellene, holder land i Europa intervjuer i mellom 80 % og 100 % av sine personell seleksjonsprosesser (Dany & Torchy, 1994). Det er nok mange grunner til denne populariteten, blant annet den høye aksepten for intervjuet

(Lievens, 2007), fordi metoden lett lar seg tilpasse, fordi organisasjonen og kandidaten kan presentere seg selv, og fordi det føles intuitivt riktig for mange å selektere basert på egen evne til å bedømme mennesker. Intervjuet er også vist å være den seleksjonsmetoden som kandidatene liker best (Hausknecht, Day, & Thomas, 2004).

Til tross for den høye populariteten har ikke intervjuet den samme sterke sammenhengen med jobbprestasjoner som for eksempel GMA-tester. Kjenstadbakk (2012). Vik (2013) og Isaksen (2014) fant på lignende utvalg som denne studien, at intervjuet tilførte lite eller ingen forklart varians etter at GMA og gjennomsnitt fra videregående skole var kontrollert for. Cook (2009) oppsummerte fire meta-analyser av Hunter og Hunter (1984) (30 studier), Wiesner og Cronshaw (1988) (160 studier), Huffcutt og Arthur (1994) (114 studier) og McDaniel, Whetzel, Schmidt og Maurer (1994) (245 studier) og viste at gjennomsnittsvaliditeten for ustrukturerte intervjuer med jobbprestasjoner som kriterium, var $r = .15$ ukorrigert, og for strukturerte intervju økte denne til $r = .28$. Selv om Sjøforsvaret benytter seg av intervjuguider og intervjukurs, er intervjuet som gjennomføres å betrakte som et relativt ustrukturert intervju, snarere enn et strukturert. Dette er blant annet fordi guiden fungerer som et støtteelement heller enn noe ufravikelig, og fordi det ikke er noen strukturert måte å vurdere om svaret kandidaten gir er dårlig, middels eller bra. Nyere reanalyser av McDaniel et al. (1994) fant for øvrig høy grad av publikasjonsbias for strukturerte intervju (Duval, 2005). Publikasjonsbias innebærer at publiserte artikler ikke er representative for alle studier som er gjennomført på et område og typisk har sterkere funn enn de upubliserte studiene, enten fordi forskere ikke forsøker å publisere svake og ikke-signifikante resultater, eller fordi journalene ikke synes funnene er spennende nok. Da denne feilkilden ble tatt høyde for var det i denne meta-analysen ingen forskjell i validitet mellom strukturerte og ustrukturerte intervjuer. Dette antyder at selv om strukturerte intervju har mye støtte for å være mer valide enn ustrukturerte intervju, er ikke debatten ennå helt avgjort (Dalton,

Aguinis, Dalton, Bosco, & Pierce, 2012). Strukturerte intervju viser å ha andre fordeler knyttet til seleksjon. Terpstra, Mohamed og Kethley (1999) fant at strukturerte intervju ble klaget på halvparten så ofte som forventet og ustrukturerte intervju ble klaget på dobbelt så ofte som forventet. I tillegg til dette ble det funnet at, av de klagene som endte i retten, vant de som hadde brukt strukturert intervju nesten alle mens de som hadde brukt ustrukturert intervju vant kun 60 % av sakene.

Vurderingssenter som seleksjonsmetode

Et vurderingssenter (assessment center) består av en standardisert evaluering av adferd basert på flere kilder (Rupp et al., 2015). Flere trente observatører blir vanligvis brukt til å vurdere jobbsimuleringer spesifikt designet for stillingen. Disse vurderingene blir deretter samlet og nivellert ved diskusjon eller en statistisk integreringsmetode. Dette resulterer i evalueringer av kandidatenes prestasjoner innenfor de dimensjonene jobbsimuleringen var ment å måle. Det er viktig at vurderingssenteret blir utviklet, implementert og validert for den spesifikke stillingen (eller de spesifikke stillingene) den skal brukes til å selektere til. Hvis man baserer seg på blindt på metoder utviklet andre steder, selv om disse stillingene virker relativt like, risikerer man bruke metoder som ikke lar seg generalisere. Hvis man lar seg inspirere av andre vurderingssenter er det derfor kritisk at metoden både pre-testes på lignende eller samme populasjon og valideres i ettertid på samme populasjon.

Det er ifølge «International Task Force on Assessment Center Guidelines» (Rupp et al., 2015) ti elementer som må være på plass for at man skal kunne kalle en metode for et vurderingssenter. 1. Først og fremst må det være en jobbanalyse eller kompetansemodell i bunn som leder til klart definerte kompetanser. De kompetansene som skal testes for i vurderingssenteret må også kunne observeres som adferd og være knyttet til prestasjon innenfor relevante felt. 2. Den observerte adferden må kunne bli korrekt klassifisert inn i sin kompetanse av trente observatører. 3. Det må være flere komponenter (hvis det bare er en

komponent, kalles metoden simulering eller arbeidscase), og hver av disse komponentene må være pre-testet for å sikre at de bidrar med reliabel, objektiv og relevant informasjon. 4. Det må eksistere en oversikt over hvilke kompetanser som måles i hvilke komponenter (en dimensjon-øvelse matrise). 5. Hver adferd må kunne bli observert flere ganger gjennom jobbrelevante simuleringer og trente observatører må kunne se den samme adferden (disse simuleringene bør også ligne den faktiske jobben mest mulig, i form av oppgave, ikke nødvendigvis setting). 6. Flere observatører må bli brukt til å observere samme kandidat og det bør være et begrenset antall kandidater som observeres om gangen. 7. Alle som skal observere må få grundig gjennomgang i hva oppgaven går ut på, og de må kunne demonstrere at de mestrer oppgaven i tilstrekkelig grad. 8. Adferden må kunne loggføres på en systematisk og korrekt måte samtidig som adferden skjer, eller ved hjelp av video-/lydopptak. 9. Integreringen av observasjonene av hver kandidat må være basert på en diskusjon der alle observasjonene ligger til grunn, eller ved bruk av en statistisk integreringsmetode. Det er viktig at informasjon samlet utenfor vurderingssenteret ikke brukes i denne sammenheng. 10. Vurderingssenteret må være standardisert slik at alle kandidatene får lik mulighet til å demonstrere sine ferdigheter, egenskaper og evner. Dette inkluderer instruksjoner gitt, tid til disposisjon, tilgjengelig utstyr, fasiliteter, adferden til rollespillere, rekkefølge av oppgaver m.m. Disse 10 retningslinjene er utviklet for å sikre en viss standardisering av metoden. Et tiltak som er ment å utnytte metodens potensiale, slik at validiteten øker.

Vurderingssenter er, som intervjuet, en metode som tilpasses organisasjonen i større grad enn for eksempel tester som måler GMA. Dette er langt på veg en styrke for metoden, men kan fort bli en svakhet hvis organisasjonen bruker komponenter som ikke er av tilstrekkelig kvalitet, eller komponenter som ikke lar seg generalisere til dette tilfellet. I tre meta-analyser ble det funnet at kriterievaliditeten mellom vurderingssenteret og jobbprestasjon (målt med sjefens mening) var omtrent .23 og den korrigerte validiteten var

omtrent .30 (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2007; Hermelin, Lievens, & Robertson, 2007), noe som også virker å gjelde for militærspesifikke analyser (Damitz, Manzey, Kleinmann, & Severin, 2003; Dobson & Williams, 1989; Isaksen, 2014; Kjenstadbakk, 2012; Melchers & Annen, 2010). Dette kan virke som relativt lav korrelasjon med tanke på at vurderingssenteret som regel er en svært ressursekrevene seleksjonsmetode, men det er sannsynlig at denne er noe lavere enn den reelle validiteten. Det er fordi vurderingssenter er svært ressurskrevende og blir brukt til slutt i en seleksjonsprosess, noe som sannsynlig fører til at spredningen er ytterligere begrenset utover det man tar høyde for i en korrigering av validiteten (Hardison & Sackett, 2007). Likevel kan man argumentere for at validering av vurderingssenter som helhet ikke er nyttig, nettopp fordi et vurderingssenter kan være veldig forskjellig fra et annet. En annen måte å løse valideringsspørsmålet er derfor å finne validiteten til hver enkelt kompetanse. Dette er en lønnsom strategi fordi ikke alle kompetanser er like relevante opp mot jobbprestasjon eller like lette å måle (Arthur, Day, McNelly, & Edens, 2003), og fordi noen kompetansemål kan være overflødige og ikke bidra til en økning i validitet (Feltham, 1988). Forskning gjort på det Israelske sjøforsvaret fant at jobbsimuleringer på entring av gummibåt, navigering og teltoppsett predikerte prestasjoner ett år senere (Rom & Kalderon, 2013). Når det gjelder tilleggssvalidteten til vurderingssenteret etter GMA-tester var kontrollert for, fant forskning gjort på tysk og israelsk politi at denne var høy, og de anbefaler derfor å bruke disse metodene sammen (Dayan, Kasten, & Fox, 2002; Krause, Kersting, Heggestad, & Thornton, 2006).

Selv om vurderingssenteret er regnet som en valid metode, er det noen klare metodiske begrensninger man bør ta stilling til. Siden kandidatene arbeider i grupper, vil prestasjonene til én være avhengig av hvem som er i samme gruppe. En kandidat som er over gjennomsnittlig dominant, vil kunne virke tilbaketrukken i en gruppe med ekstremt dominante mennesker. Dette problemet kan reduseres ved å sette sammen gruppene nøye eller designe

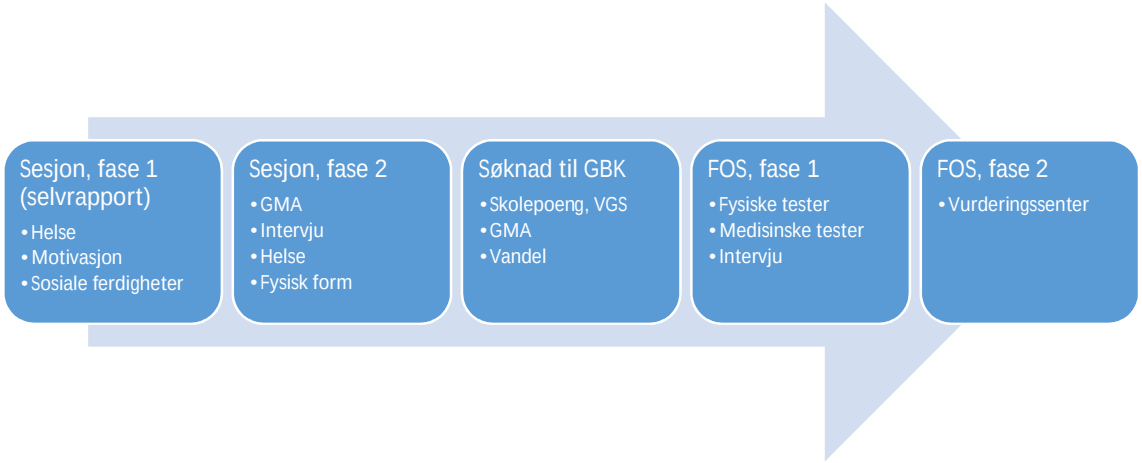
simuleringer med rollespillere som oppfører seg likt hver gang, men det vil alltid være et aspekt man bør ta hensyn til. Et annet potensielt problem kan være at de som presterer godt i vurderingssenteret blir behandlet på en annen måte når de kommer inn på utdanningen eller inn i stillingen. Dette kan forurense kriteriet ved å skape en selvoppfyllende profeti (Merton, 1948) eller glorie-effekt (Nisbett & Wilson, 1977) og dermed en falsk validitet. Det er derfor svært uheldig hvis arbeidsgiver/utdanningsansvarlig får innsyn i hvordan de nytilsatte presterte på seleksjonsarenaen - eller enda verre, hvis samme person som bedømte prestasjonen på vurderingssenteret også vurderer prestasjonen i jobben. Når det gjelder indirekte diskriminering er dette funnet å være et problem også for vurderingssenter. Dean et al. (2008) fant i sin meta-analyse at etniske minoriteter blir valgt sjeldnere enn hva man skulle anta (d = .52 for de med afrikansk bakgrunn og d = .40 for de med sør-amerikansk bakgrunn). I tillegg til dette ble det også funnet en svak indirekte diskriminering mot menn (d = .18).

Felles opptak og seleksjon

«Felles opptak og seleksjon» (FOS) er en omfattende seleksjonsarena som holdes to ganger i året, en om vinteren og en om sommeren. Omtrent 5000 unge kvinner og menn søker hvert opptak på forskjellige militære utdanninger, og blant disse er Sjøforsvarets grunnleggende befalskurs (GBK). Intensjonen med FOS er å selektere de kandidatene som har størst sjanse, blant disse 5000 interesserte, til å prestere godt innenfor den utdanning de har søkt seg til. Opptaket varer uavbrutt i to uker, og kandidatene blir selektert basert på medisinske tester, fysiske tester, psykologiske tester, intervju og et syv dager langt vurderingssenter. Informasjonen som samles inn på FOS er ment å si noe om kandidatenes potensiale til å prestere i de teoretiske emnene de møter under utdanningen og kandidatenes potensiale som militær leder. For å systematisere dette blir de som består alle testene gitt en skoleprognose, en lederprognose og en offisersvurdering, hvor en vektet sum av disse er

konkurransepoengene som bestemmer rangeringen av kandidatene og hvem som kommer inn på utdanningen de har søkt.

I tillegg til dette har alle søkerne allerede gjennomført sesjon del 2, hvor alle 18 år gamle kvinner og menn som ved selvrapportering fyller Forsvarets krav til helse, motivasjon og sosiale ferdigheter (sesjon del 1), blir kalt inn. Her grovselekterer man basert på fysisk form, intervju (en offisersvurdering av motivasjon og egnethet), GMA, medisinske tester, vandel og prestasjoner fra videregående skole. Seleksjonen på sesjon reduserer gruppen til omtrent 15 % av sin opprinnelige størrelse, fra omtrent 60 000 som besvarer sesjon del 1 til omtrent 9000 som blir fordelt til førstegangstjeneste på sesjon del 2 (Køber, 2015).



Figur 3: Seleksjonstrinn før opptak til Sjøforsvarets GBK

Selv om FOS er ansett som et svært besparende tiltak har denne arenaen betraktelige kostnader knyttet til seg, både økonomisk og fordi opptaket er en stor påkjenning på deltagerne. Det er på grunn av disse kostnadene på organisasjonen og kandidatene at det er viktig at alle testene bidrar til å predikere fremtidige resultater. Det er til nå ikke gjennomført valideringsstudier som undersøker hvorvidt FOS sine opptakstester faktisk predikerer resultatene på GBK, bare hvorvidt det predikerer resultatene på andre befalsutdanninger (Isaksen, 2014; Kjenstadbakk, 2012). Målet med denne studien er derfor å kvalitetssikre

opptaket til Sjøforsvarets GBK, slik at Forsvaret med større sikkerhet kan si om FOS, slik det er designet i dag, fungerer til å velge ut befal til denne utdanningen.

Sjøforsvarets grunnleggende befalskurs

Sjøforsvarets GBK er en fire måneder lang grunnleggende militær lederutdanning. Det overordnede målet med utdanningen er å gi elevene de beste forutsetningene for å besette lederfunksjoner og spesialistfunksjoner på lavere avdelingsnivå. Utdanningen fokuserer derfor på å utvikle kunnskaper, ferdigheter og holdninger som er avgjørende for å nå dette målet. Eksempler på dette er: utdanning og trening av underlagt personell, kommando og ledelse av egen enhet, ta ansvar og vise omsorg for sitt personell og vise initiativ og handlekraft for å løse oppdrag i tråd med sjefens intensjon. For å oppnå dette deles utdanningen inn i fire kjerneområder: 1. Ledelse: Består av militær lederskap, pedagogikk, lover og etikk, 2. Militær idrett og trening, 3. Stridsmiljø og overlevelse: Inneholder grunnleggende soldatferdigheter (våpentjeneste, militære øvelser m.m.). 4. Militære profesjonsfag: Inneholder stabs –og driftsfag (sikkerhetsstyring og risikovurdering, forsvarskunnskap, forvaltningstjeneste m.m.). Disse kjerneområde testes både teoretisk og praktisk gjennom hele utdanningsperioden. I grove trekk er utdanningen inndelt i en to uker lang grunnpakke, etterfulgt av ni uker på båt, tre uker med infanteritjeneste og en uke med oppsummering og eksamen.

Problemstillinger i denne studien

Formålet med denne studien er å undersøke den prediktive validiteten til GMA, skolepoeng, intervjuet og vurderingssenteret opp mot prestasjonene på Sjøforsvarets GBK. Prestasjonene på GBK er delt inn i variabelen «Hovedkarakter», som er summen av alle tellende emner, og «Skikkethet som militær leder», som er et mål på hvor egnet eleven er som militær leder. Hvis det viser seg at det ikke er en tydelig kobling mellom prediktorene og prestasjonene etter endt utdanning må man spørre seg selv hvorvidt det er etisk og økonomisk

forsvarlig å gjennomføre et så omfattende og krevende opptak. Videre vil det bli sett på hvilken grad av forklart varians hver test på FOS tillegger etter at de eksisterende testene er kontrollert for. For intervjuet er dette GMA og gjennomsnittspoeng for videregående skole, og for vurderingssenteret kontrolleres det også for intervjuet. Studien er designet slik fordi de siste testene er de som koster mest, både fra kandidatenes side og Forsvarets side, og bør derfor bidra utover de andre testene. Dette gir de følgende problemstillingene:

Er det en positiv korrelasjon mellom prediktorene og prestasjoner på Sjøforsvarets GBK?

Er det slik at alle prediktorene tilfører forklart varians av prestasjonene på Sjøforsvarets GBK etter at de foregående prediktorene er kontrollert for?

Det er forventet at «Skolepoeng, VGS», «Alminnelig evnenivå» og «Skoleprognose» kommer til å være de sterkeste prediktorene for «Hovedkarakter», og at «Offisersvurdering» og «Lederprognose» kommer til å være de sterkeste prediktorene for «Skikkethet som militær leder». Denne forventningen er basert på en tematisk overflatevaliditet mellom disse variablene. I tillegg til dette er det forventet at den prediktive validiteten til intervjuet og den totale forklarte variansen opp mot «Skikkethet som militær leder» vil være lav. Denne forventningen er basert på studier gjort på lignende populasjoner i det norske Forsvaret (Kjenstadbakk, 2012; Isaksen, 2014).

Metode

Utvalg

Utvalget består av de elevene som har gjennomført hele opptaket på FOS, og som har blitt selektert inn til Sjøforsvarets GBK og gjennomført denne utdanningen i perioden våren 2010 – høsten 2014. Noen av elevene gikk direkte til GBK etter endt opptak, mens andre søkte om utsettelse og tok utdanningen senere (maks ett år). De som fikk søke på denne

utdanningen var allerede i en militær enhet hvor de hadde vist potensiale som militært befal og dermed fått formell anbefaling av sin overordnede om å søke på opptaket. Elevene som kom inn på utdanningen hadde dermed alle militær erfaring, men denne varierte fra påbegynt førstegangstjeneste til flere år med operativ erfaring. Elevene hadde alle søkt på denne utdanningen først etter at arbeidsgiver hadde vist til en konkret plan og stilling for vedkommende.

Totalt 444 elever begynte utdanningen sin på GBK i Sjøforsvaret i denne perioden. Av disse sluttet 11 stykker før hoveddelen av eksamener kunne bli gjennomført, og ble dermed ekskludert fra analysene. Studiens totale antall deltagere er dermed 433, noe som tilsvarer 97.5 % av de som startet GBK. Av disse 433 var 403 elever menn (93.1 %) og 30 elever kvinner (6.9 %). Elevene var mellom 19 og 30 år.

Måleinstrumenter og prosedyre

I studien innhentes data på tre forskjellige tidspunkt; før opptaket (T1), under opptaket (T2) og etter endt utdanning (T3). Studiens mange variabler fordelt på tre tidsperioder medførte at det mangler datapunkt hos flere av elevene. Det endelige datamaterialet besto av cirka 290 personer for T2-testene og cirka 430 personer for variablene «Alminnelig evnenivå» og «Hovedkarakter». Variablene med størst grad av «missing» var «Skolepoeng, VGS» (N = 187) innsamlet på T1 og «Skikkethet som militær leder» (N = 172) innsamlet på T3.

Før opptaket (T1):

«Alminnelig evnenivå» (GMA) ble målt på sesjon da elevene var mellom 18 og 19 år. Dette gjaldt i perioden før 2009 alle gutter i hvert årskull, samt de jentene som takket ja til sesjon. Etter 2009 fikk jenter sesjonsplikt og færre gutter ble kalt inn på sesjon del 2 der «Alminnelig evnenivå» ble målt. Elevene i denne studien havnet også midt i endringen fra pen og papir til PC, en endring som har vist seg å ikke ha noen signifikant påvirkning på

egenskapene til testene (Skoglund, Martinussen, & Lang-Ree, 2014). Med unntak av dette ble testene likt administrert for alle når det gjaldt instruksjon og tidsbruk. Ved testing av «Alminnelig evnenivå» administreres tre deltester med tidsbegrensning: Regneoppgaver (U4), Figurregler (U5) og Ordlikhet (U6). U4 måler matematiske ferdigheter og resonneringsevne gjennom 30 praktiske regneoppgaver og varer i 25 minutter, U5 måler abstrakt resonneringsevne gjennom 36 oppgaver og varer i 25 minutter, og U6 måler ordforståelse gjennom 54 oppgaver og varer i seks minutter (Skoglund et al., 2014). Alle deltestene besvares med «multiple-choice» og skåres som antall riktige svar. Antall riktige svar regnes så om til persentiler. «Alminnelig evnenivå» regnes ut for hver elev som summen av persentilene for disse tre deltestene. Test-retest reliabiliteten er beregnet i tidligere studier og er for U4 = .84, U5 = .72 og U6 = .90 (Sundet, Tambs, Magnus, & Berg, 1988).

Skolepoeng fra videregående skole er hentet fra skoledokumentene som elevene sendte inn da de søkte på GBK i forkant av opptaket. Den endelige poengsummen er regnet som summen av alle tellende emner fra videregående skole uten tilleggs-poeng. Denne poengsummen ble så ført inn i Forsvarets personalsystem «P3», en prosedyre som i denne perioden bare delvis ble fulgt noe som førte til en høy andel «missing» data (se Tabell 1). Det virker ikke å være en systematikk i når det er blitt ført og når det ikke er blitt ført.

Under opptaket (T2):

Intervjuene ble gjennomført av to offiserer som sammen satt en lederprognose og en skoleprognose basert på et 30 minutter langt intervju (15 minutter forberedelse, 30 min intervju og 15 min etterarbeid). Hver prognose ble angitt på en ni-delt skala og elevene måtte skåre tre eller høyere på hver prognose for å komme videre i opptaket. Momentene som ble vektlagt i intervjuet var for skoleprognosen; motivasjon, studievaner, fysisk robusthet og konkurranse –eller prestasjonsorientering. For lederprognosen var momentene; evne til å beskrive egne lederegenskaper, selvinnsikt på egne utviklingsområder, kommunikasjon og

tidligere erfaringer. I tillegg hadde offiserene informasjon om «Alminnelig evnenivå», «Skolepoeng, VGS» og resultater på fysiske tester som vektet inn på skoleprognosen, og tjenesteuttalelser som vektet inn på lederprognosen. Hvis offiserene mente det var behov for et re-intervju, ofte grunnet utilstrekkelig informasjon, ble dette utført av en psykolog. For å tilstrebe god reliabilitet og lik behandling av søkerne, fikk panelene en intervjuguide, et skåringsark og et intervjukurs som sammen skulle hjelpe dem til å komme frem til prognosene. Prognosene som ble satt ble ført inn i Forsvarets journalsystem «P3».

Offisersvurderingen ble gjort i løpet av en syv dager lang militær feltøvelse hvor elevene deltok i flere realistiske oppgaver hvor de byttet på å være leder. En nøye utvalgt offiser fulgte sitt respektive lag gjennom hele denne perioden. Øvelsen begynte med omtrent et døgn med hard fysisk intensitet, etterfulgt av tre dager med jobbsimuleringer fra det militære yrket, og den siste delen brukte hver offiser slik han/hun ønsket for å belyse de personene og kompetansene hvor det ennå var tvil. I løpet av denne perioden var elevene delt inn i lag på 8-10 personer, og de bodde og samarbeidet med denne gruppen 24 timer i døgnet. Mot slutten av feltøvelsen satt offiseren som hadde fulgt det respektive laget, en prognose mellom 1-9 for hver elev. Denne prognosen skulle gjenspeile elevens evne til å ta initiativ, håndtere usikkerhet, skape tillit, vise omsorg og evne til å ta selvstendige og gode beslutninger. Eleven måtte ha minst karakter tre for å bli vurdert til GBK. For å forberede og nivellere offiserene ble det gitt kurs som inkluderte teori og praktiske øvelser. Nivellering skjedde også kontinuerlig i løpet av øvelsen. Den endelige prognosen ble ført inn i Forsvarets journalsystem «P3».

Etter endt utdanning (T3):

«*Hovedkarakter*» er den endelige karakteren etter endt GBK. Denne karakteren gjenspeiler et vektet gjennomsnitt av alle tellende emner. De tellende emnene er Ledelsesfag (vektet 16 poeng), etikkfag (vektet 4 poeng), undervisningslære (vektet 10 poeng), Idrett og

trening (vektet 10 poeng), Militær profesjon (vektet 10 poeng), Stridsteknikk (vektet 10 poeng), Stridsmiljø og overlevelse (vektet 10 poeng), Holdningsfag (vektet 10 poeng) og Sjømannskap (vektet 10 poeng). Denne vektingen har endret seg i løpet av perioden ved at noen fag har gått fra å vekte til å bli målt som bestått/ikke-bestått. Noen av fagene er praktisk orientert (hhv. Undervisningslære, Sjømannskap og Stridsmiljø og overlevelse), noen er teoretisk orientert (hhv. ledelsesfag, etikkfag, militær profesjon og holdningsfag) og noen er blandet (hhv. Idrett og trening). Karakteren i hvert emne angis fra 1-6. Variabelen «Hovedkarakter» rapporteres med to desimaler for å gi et mer nyansert bilde.

«Skikkethet som militær leder» (SML), tidligere kjent som «Militært forhold» (MIFO), er en variabel som måler hvor egnet eleven er som militær leder. Denne karakteren favner i praksis alt eleven gjør i løpet av sin utdanning og vurderer hvorvidt eleven har utviklet egen lederrolle, lederegenskaper, selvforståelse, mestringstro og holdninger i tråd med Sjøforsvarets intensjon. Emnets endelige karakter er nærmeste overordnende sin vurdering av dette, en offiser som har fulgt eleven tett i fire måneder. Denne vurderingen fylles inn i en strukturert tjenesteuttalelse. SML måles i fem ledd fra under norm til over norm og er kodet om på en 1-5 skala, men vektes ikke inn i hovedkarakteren siden det er tilstrekkelig at eleven består emnet. SML-karakteren har data fra kun tre av kullene, grunnet manglende lagring av den graderte informasjonen fra de andre kullene.

Endringer i datasettet

Karakterskalaen gikk fra 1-6 til 1-9 sommeren 2012 for variablene «Skoleprognose», «Lederprognose» og «Offisersvurdering». Hvis den gamle skalaen ble brukt ved flere anledninger hvor 1-9 skulle blitt brukt, kan dette ha resultert i en systematisk feil. En svært nøye gjennomgang av datamaterialet ble gjort for å unngå denne problematikken, så det er forventet at dette gjelder svært få. De som ble funnet ført 1-6 ble regnet om til 1-9 slik at alle ble vurdert på samme skala.

Det har skjedd noen endringer i fagene på GBK mellom 2010 og 2014. Disse endringene består i hovedsak av fag som ble slått sammen. Endringene er minimale, og det forventes ikke at de har hatt noen betydning for vurderingen av kandidatene og den gjennomsnittskarakteren de oppnådde.

Etikk

NSD (Norsk samfunnsvitenskapelig datatjeneste) anslår studien som ikke-meldepliktig. Dette fordi forskeren ikke mottok data som kunne direkte eller indirekte identifisere enkeltpersoner. Det ble heller ikke gitt en koblingsnøkkel mellom de identifiserbare og ikke-identifiserbare datasettene. Anonymiseringen av datasettet ble gjort av Sjefpsykologen for Forsvaret.

Statistiske analyser

IBM SPSS Statistics 21 ble brukt til å gjennomføre alle de statistiske analysene. Korrelasjonsanalyse ble brukt i denne studien for å utforske sammenhengen mellom de ulike testene og prestasjonene oppnådd etter fire måneder med befalsskole. Hierarkisk multippel regresjonsanalyse ble benyttet for å se hvor mye av variansen i prestasjonene på GBK prognosene fra intervjuet på FOS forklarte etter at «Alminnelig evnenivå» og «Skolepoeng, VGS» var kontrollert for, og hvor mye av variansen «Offisersvurdering» på FOS forklarte det samme når «Alminnelig evnenivå», «Skolepoeng, VGS» og prognosene fra intervjuet var kontrollert for. Hierarkisk, trinnvis regresjon viser også det relative bidraget til hver variabel. Antall elever varierer fra 91 til 433 i de ulike analysene. For å tolke størrelsen på korrelasjonene ble kategoriseringen liten, medium og sterk basert på Cohens retningslinjer brukt (Cohen, 1988).

De uavhengige variablene (prediktorer) var «Alminnelig evnenivå», «Skolepoeng, VGS», «Skoleprognose», «Lederprognose» og «Offisersvurdering» (T1 + T2), og de avhengige variablene (kriterier) var «Hovedkarakter» og «Skikkethet som militær leder» (T3).

Før korrelasjonsanalysen og regresjonsanalysen ble gjennomført var det ønskelig å se etter hvorvidt «scatterplotet» viste brudd på antagelsen om linearitet, homogenitet av variansen («homoscedasticity») eller hvorvidt det var noen ekstreme utliggere. Statistiske utliggere ble også testet for, med «Mahalanobis distance» og «Cooks distance». Videre ble brudd på antakelsene om normalitet og multikolinearitet vurdert ved hjelp av hhv. histogram og «collinearity diagnostics». Det ble også vurdert hvorvidt det var for høy «missing» data for variablene «Skolepoeng, VGS» og «Skikkethet som militær leder», og hvorvidt hver av variablene oppfylte antagelsene om uavhengighet mellom observasjoner (Stevens, 1996). En vurdering av hvorvidt analysene holdt tilstrekkelig statistisk styrke ifølge eksisterende teori ble også gjort (Tabachnick & Fidell, 2013).

Resultater

Innledende analyser av datamaterialet viste ingen brudd på antagelsene om linearitet, homogenitet av variansen, statistiske utliggere, multikolinearitet eller manglende normalitet. Det ble funnet at utvalget var tilstrekkelig stort for alle analysene (Tabachnick & Fidell, 2013). Formelen som ble benyttet for regresjonsanalysene var $N > 50 + 8m$, hvor m er antall uavhengige variabler. For regresjonsanalysene betyr dette at andelen personer bør overstige 90 i hver analyse.

Det ble funnet brudd på antagelsen om uavhengighet mellom observasjoner for variablene «Offisersvurdering» og «Skikkethet som militær leder». Disse variablene baserte seg på prestasjoner i grupper og var derfor til dels avhengig av hvordan de andre gruppemedlemmene presterte (Stevens, 1996). I denne oppgaven ble det vurdert tilstrekkelig

med en teoretisk kompensasjonen i form av kritikk og diskusjon. Det eksisterer heller ikke data på hvem som deltok i hvilke grupper, så det er ikke mulig å foreta noen analyser der dette modelleres.

Deskriptiv statistikk

Tabell 1 synliggjør de varierende måtene hver variabel skåres på, samt det varierende antallet datapunkt («*N*»). Tabellen viser en varierende grad av manglende data («missing»), hvor «Skikkethet som militær leder» og «Skolepoeng, VGS» mangler mest med hhv. 60.3 % og 56.8 %.

Tabell 1

Deskriptiv statistikk for variablene i studien

	Minimum	Maksimum	N	Missing
Alminnelig evnenivå	2	9	426	1.6 %
Skolepoeng, VGS	26.40	55.00	187	56.8 %
Skoleprognose	2	9	287	33.7 %
Lederprognose	2	9	286	33.9 %
Offisersvurdering	2	9	298	31.2 %
SML	2	5	172	60.3 %
Hovedkarakter	1.19	4.60	433	0 %

Note: «Skolepoeng, VGS» = karaktersum fra videregående skole, SML = «Skikkethet som militær leder»

Sammenheng mellom de inkluderte variablene

Deskriptiv statistikk og korrelasjoner mellom variablene er presentert i Tabell 2. Alle prediktorene var positivt korrelert med «Hovedkarakter». De sterkeste korrelasjonene ble beregnet for «Skolepoeng, VGS» (r = .39), «Skoleprognose» (r = .33) og «Alminnelig evnenivå» (r =. 31) som alle var middels høyt korrelert med «Hovedkarakter». «Offisersvurdering» (r = .25) og «Lederprognose» (r = .24) litt svakere korrelert med «Hovedkarakter» (Cohen, 1988).

Kriteriet «Skikkethet som militær leder» var også positivt korrelert med alle prediktorene, unntatt «Alminnelig evnenivå» der korrelasjonen ikke var statistisk signifikant (r = .03). De sterkeste korrelasjonene ble beregnet for «Offisersvurdering» (r = .41) og «Lederprognose» (r = .40). «Skoleprognose» (r = .22) og «Skolepoeng, VGS» (r = .29) var svakt korrelert med denne variabelen (Cohen, 1988).

Tabell 2

Gjennomsnitt, standardavvik og korrelasjoner mellom variablene i studien.

	M (SD)	1	2	3	4	5	6
1. AE	5.56 (1.33)	-					
2. Skolepoeng, VGS	40.32 (6.54)	.30**	-				
3. Skoleprognose	5.95 (1.38)	.38**	.58**	-			
4. Lederprognose	6.09 (1.58)	.14*	.14	.42**	-		
5. Offisersvurdering	6.14 (1.61)	.10	.14	.14*	.49**	-	
6. SML	3.34 (0.74)	.03	.29**	.22*	.40**	.41*	-
7. Hovedkarakter	2.89 (0.60)	.31**	.39**	.33*	.24**	.25**	.47**

Note: SML= Skikkethet som militær leder, AE = Alminnelig evnenivå, Skolepoeng, VGS = karaktersum fra videregående skole.

*p < .05, **p < .01 (two-tailed).

Hierarkisk multippel regresjonsanalyse

To hierarkisk multiple regresjonsanalyser ble gjennomført, hvor den eneste forskjellen var den avhengige variabelen. I den første analysen ble «Hovedkarakter» brukt som avhengig variabel og i den andre analysen ble «Skikkethet som militær leder» brukt som avhengig variabel. Begge analysene ble gjort med «exclude cases pairwise» grunnet høy «missing» for variablene «Skikkethet som militær leder» og «Skolepoeng, VGS». Antallet deltagerne i analysene var $N = 145$ med «Hovedkarakter» som avhengig variabel og $N = 91$ med «Skikkethet som militær leder» som avhengig variabel.

«Hovedkarakter»

«Alminnelig evnenivå» og «Skolepoeng, VGS» ble lagt inn i trinn 1 og forklarte til sammen 19 % av variansen i «Hovedkarakter». I trinn 2 ble intervjuets «Skoleprognose» og «Lederprognose» lagt til, noe som ikke resulterte i en signifikant økning i forklart varians. Trinn 3 tilførte «Offisersvurderingen», noe som heller ikke ga en signifikant økning i forklart varians. I den siste modellen var bare variablene fra trinn 1 statistisk signifikante, hvor «Skolepoeng, VGS» hadde $\beta = .27$ og «Alminnelig evnenivå $\beta = .19$. Samlet sett forklarte modellen 24 % av variansen i kriteriet «Hovedkarakter».

Tabell 3

Hierarkisk regresjonsanalyse av hvordan opptakstestene predikerer den samlede hovedkarakteren ved endt utdanning.

	Trinn 1	Trinn 2	Trinn 3
	β	β	β
Alminnelig evnenivå	.22**	.19*	.19*
Skolepoeng, VGS	.32**	.29**	.27**
Skoleprognose		.02	.06

Lederprognose		.16	.08
Offisersvurdering			.15
R^2 total	.19**	.22**	.24**
R^2 Change	.19**	.03	.02

Note: *$p < .05$, **$p <. 01$, $N = 145$

«Skikkethet som militær leder»

«Alminnelig evnenivå» og «Skolepoeng, VGS» ble lagt inn i trinn 1 og forklarte til sammen 9 % av variansen i «Skikkethet som militær leder». I trinn 2 ble intervjuets «Skoleprognose» og «Lederprognose» lagt til, noe som økte den totale andel forklart varians til 23 %. Trinn 3 tilførte «Offisersvurderingen» som økte den totale andel forklart varians til 28 %. Intervjuet gitt under FOS økte altså den totale andel forklart varians med 14 % etter at «Alminnelig evnenivå» og «Skolepoeng, VGS» var kontrollert for, og «Offiservurderingen» økte den totale forklaringen av varians med ytterligere 5 % etter at «Alminnelig evnenivå», «Skolepoeng, VGS» og intervjuet var kontrollert for. Begge disse økningene var statistisk signifikante. I den siste modellen var «Skolepoeng, VGS», «Lederprognose» og «Offisersvurdering» statistisk signifikante, og hadde omtrent samme betaverdi (β = .26). Samlet sett forklarte modellen 28 % av variansen i kriteriet «Skikkethet som militær leder».

Tabell 4

Hierarkisk regresjonsanalyse av hvordan opptakstestene predikerer karakteren «Skikkethet som militær leder» ved endt utdanning.

	Trinn 1	Trinn 2	Trinn 3
	β	β	β
Alminnelig evnenivå	-.06	-.08	-.10

Skolepoeng, VGS	.31**	.31**	.27*
Skoleprognose		-.10	-.04
Lederprognose		.41**	.26*
Offisersvurdering			.26*
R^2 total	.09*	.23**	.28**
R^2 Change	.09*	.14**	.05*

Note: *p <. 05, **p < .01, N = 91

Diskusjon

Litteraturen på feltet har vist at alle prediktorene som benyttes i denne analysen er forventet å korrelere med jobbprestasjoner på tvers av yrker. Dette gjelder gjennomsnittlig karakter fra videregående skole (Isaksen, 2014; Kjenstadbakk, 2012; Roth et al., 1996; Vik, 2013), generelle mentale evner (Hunter & Hunter, 1984; Judge et al., 1999; Kjenstadbakk, 2012; Melchers & Annen, 2010; Murphy, 2002; Ree & Earles, 1992; Schmidt, 2002; Schmidt & Hunter, 1998, 2004; Sundet et al., 2004), intervju (Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; McDaniel et al., 1994; Wiesner & Cronshaw, 1988) og vurderingssenter (Damitz et al., 2003; Dobson & Williams, 1989; Gaugler et al., 1987; Hardison & Sackett, 2007; Hermelin et al., 2007; Isaksen, 2014; Kjenstadbakk, 2012; Melchers & Annen, 2010). Målet med denne studien var å finne ut hvorvidt disse korrelasjonene også gjaldt for Sjøforsvarets grunnleggende befalskurs. Det var også aktuelt for å se det relative bidraget til hver av disse prediktorene, samt hvorvidt testene tilførte noe nytt når GMA, gjennomsnittspoeng fra videregående skole var tatt høyde for (for vurderingssenteret ble også intervjuet tatt høyde for). Forventningen basert på litteraturen var å finne at vurderingssentret ville gi økt forklart varians utover bidraget fra GMA-testene (Dayan et al., 2002; Krause et al.,

2006) og at økningen ville være lav for intervjuet (Roth & Huffcutt, 2013). Funnene viste generelt at prediktorene bidrar statistisk signifikant, enten opp mot «hovedkarakter», opp mot «skikkethet som militær leder», eller opp mot begge. Spesielt uforventet var den høye forklarte variansen lederprognosen fra intervjuet og offisersvurderingen fra vurderingssenteret hadde opp mot «skikkethet som militær leder», selv etter at gjennomsnittspoeng fra videregående skole og GMA var tatt høyde for. Regresjonsanalysene viste samlet forklart varians på 28 % av «skikkethet som militær leder» og 24 % av «hovedkarakter». For «hovedkarakter» sammenfaller denne godt med eksisterende studier på lignende utvalg (Isaksen, 2014; Kjenstadbakk, 2012), mens den forklarte variansen av «skikkethet som militær leder» er høyere i denne studien (Kjenstadbakk, 2012). Det var også uventet at skoleprognosen fra intervjuet ikke tilføyde noe ekstra etter at GMA og gjennomsnittspoeng fra videregående skole var tatt høyde for. Oppsummert antyder dette at seleksjonen av GBK-elever til Sjøforsvaret som gjøres i dag, er en effektiv måte å finne de elevene som vil mestre utdanningen best.

General mental ability sin prediktive verdi

Forsvarets GMA-mål, alminnelig evnenivå, ble funnet å predikere hovedkarakteren ved endt utdanning, noe som stemmer overens med forventningene til funnene. Stryken på korrelasjonen mellom disse to variablene er også i tråd med eksisterende forskning på feltet (Hunter & Hunter, 1984; Judge et al., 1999; Kjenstadbakk, 2012; Melchers & Annen, 2010; Schmidt, 2002: Schmidt & Hunter, 1998, 2004). Disse sterke resultatene er med på å begrunne for at det settes et minstekrav til alminnelig evnenivå for denne utdanningen. Det er verdt å nevne at studien ikke sier noe om alminnelig evnenivå sin sammenheng med prestasjoner i jobb, kun opp mot prestasjonene under utdanningen. Det hadde vært interessant å se om funnene holder seg etter en slik overgang.

Det ble derimot ikke funnet sammenheng med «skikkethet som militær leder». Dette er noe overraskende, siden det er forventet at GMA skal henge sammen med jobbprestasjoner også for ledere. En potensiell forklaring er at korrelasjonen flater ut når man er tilstrekkelig intelligent. En slik forklaring antyder at minstekravet på alminnelig evnenivå for denne utdanningen i dag er tilstrekkelig høyt, hvis man kun baserer det på «skikkethet som militær leder». En annen forklaring kan være kriteriets reliabilitet, hvis denne er like lav som vist i annen litteratur (Viswesvaran et al., 1996).

Betydningen av tidligere prestasjoner

Gjennomsnittet fra videregående skole predikerte både hovedkarakter og «skikkethet som militær leder». Det er forventet at en person som presterer godt på videregående skole også vil gjøre det godt på hovedkarakteren til GBK, ettersom dette er svært like arenaer, men det er overraskende å se den samme tendensen opp mot «skikkethet som militær leder». Her ville det vært interessant å se om personlighetstrekk slik som faktoren planmessighet (conscientiousness) i fem-faktor modellen bidrar med deler av dette bildet. Det er også overraskende å se hvor sterk korrelasjonen er til tross for at det for flere av disse elevene er lenge siden de fullførte videregående skole, noe som er funnet å redusere sammenhengen mellom GPA og jobbprestasjoner (Roth et al., 1996).

Referansene brukt i denne studien (tjenesteuttalelser) er informasjon intervjueren har tilgang på under intervjuet på FOS. Basert på internasjonal forskning kan man anta at dette ikke vil være med på å øke den prediktive validiteten i særlig grad (Reilly & Chao, 1982; Hunter & Hunter, 1984), men resultatene er bedre når man ser på det militærspesifikke yrket (Jones & Harrison, 1982; Vik, 2013). Det er også vist at strukturerte referanser har god tilleggsvalidtet på GMA (Vik, 2013; Zimmerman et al., 2008), noe som kan være tilfellet for lederprognosen. Dette er med på å underbygge hypotesen om at tjenesteuttalelsen i dette tilfellet er med på å styrke intervjuets validitet. Det er vanskelig å vurdere hvor mye av

variansen intervjuet ville forklart hvis de tidligere prestasjonene ikke var tilgjengelig da vurderingen ble tatt.

Intervjuets sterke bidrag

Lederprognosen fra intervjuet korrelerer .40 med «skikkethet som militær leder» og .24 med hovedkarakter ved endt utdanning. Sett opp mot eksisterende internasjonal forskning er dette en høy korrelasjon, da forskning på ustrukturerte intervju tradisjonelt har vist betraktelig svakere funn (Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; McDaniel et al., 1994; Wiesner & Cronshaw, 1988). Dette gjelder også forskning gjort på andre utdanninger innenfor FOS-systemet, som har funnet at lederprognosen i intervjuet gir svært lav sammenheng i korrelasjonsanalysene og ingen signifikant forklart varians i regresjonsanalysen (Kjenstadbakk, 2012; Isaksen, 2014). Regresjonsanalysen i denne studien indikerer at lederprognosen har et sterkt selvstendig bidrag etter at GMA og gjennomsnittspoeng i videregående skole er tatt høyde for i forhold til prediksjon av «Skikkethet som militær leder». Grunnene til hvorfor lederprognosen er såpass sterk kan være mange. En forklaring på de sterke resultatene funnet mellom lederprognosen fra intervjuet og «skikkethet som militær leder» er bruken av tjenesteuttalelser som «anker» når intervjuet vurderes. Søkerne til GBK har ofte lenger militær erfaring enn gruppene som Kjenstadbakk (2012) og Isaksen (2014) studerte, og det er derfor større tilgang på referanser, en metode som er vist at kan bidra med moderate korrelasjoner for militære yrker (Jones & Harrison, 1982; Vik, 2013). For å utforske denne hypotesen, ville det i fremtiden vært nyttig å ta høyde for denne variabelen når GBK-opptaket skal valideres. En annen mulig forklaring er at Sjøforsvaret bruker erfarne og modne intervjuere som har jobbet sammen tidligere. Dette kan medføre en høyere validitet og reliabilitet, selv om dette heller aldri er blitt undersøkt. En annen mindre ønskelig forklaring, er at det tidvis er brukt samme person på seleksjonsarenaen som på utdanningen. Selv om dette er noe som tilhører sjeldenhetene, kan det ha vært med på

å skape en selvoppfyllende profeti som delvis forurenser validiteten (Merton, 1948). Hvorfor Sjøforsvaret lederprognose oppnår relativt sett god forklart varians av «skikkethet som militær leder", bør bekreftes i nye studier. En slik analyse vil kunne bidra med at Sjøforsvaret mer spesifikt kan fokusere på det de vet fungerer, den eksplisitte kunnskapen vil lettere kunne overføres til nye observatører og de resterende grenene kan vurdere sin praksis opp mot denne. Innen en slik analyse er gjort er det forhastet å konkludere med hvorfor resultatene er slik de er.

Skoleprognosen fra intervjuet korrelerer .22 med «skikkethet som militær leder», noe som er mer i tråd med den internasjonale litteraturen (Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; McDaniel et al., 1994; Wiesner & Cronshaw, 1988). Denne korrelasjonen øker til .33 når kriteriet byttes til hovedkarakter etter endt utdanning, noe som er å forvente, siden intervjuerne har tilgang på både GMA og gjennomsnittspoeng fra videregående skole. Interessant nok bidrar ikke skoleprognosen med forklart varians av noen av kriteriene etter at GMA og gjennomsnittspoeng fra videregående skole er tatt høyde for. Dette til tross for at intervjuet er ment å i større grad fange individet ved å ta høyde for variabler som for eksempel motivasjon og studievaner. Disse funnene indikerer at skoleprognosen kan vurderes fjernet fra intervjuet, og heller erstattes med en matematisk formel basert på GMA og gjennomsnittspoeng fra videregående skole. Dette ville enten frigjort tid til å fokusere på lederprognosen, eller frigjort ressurser ved å korte ned intervjuet. Endringer fra klinisk til matematisk prediksjon er i en meta-analyse funnet å øke den prediktive validiteten med over 50 % (Kuncel, Klieger, Connelly, & Ones, 2013).

Er vurderingssenteret i seleksjon fornuftig ressursbruk?

Prediktoren som gir «offiservurdering» omtales i denne studien som et vurderingssenter til tross for at det ikke oppfyller alle kravene til dette (Rupp et al., 2015).

Sammenligningen gjøres fordi det er store fellestrekk mellom disse, spesielt tre av dagene hvor lederegenskapene testes med jobbsimuleringer.

Vurderingssenteret korrelerer .41 med «skikkethet som militær leder» og .24 med hovedkarakteren ved endt utdanning noe som er i tråd med problemstillingen. Korrelasjonen funnet mellom vurderingssenteret og «skikkethet som militær leder» er høy sammenlignet med både internasjonal forskning (Damitz et al., 2003; Dobson & Williams, 1989; Gaugler et al., 1987; Hardison & Sackett, 2007; Hermelin et al., 2007; Melchers & Annen, 2010) og tidligere forskning på utdanninger som bruker FOS som seleksjonsarena (Kjenstadbakk, 2012; Isaksen, 2014). Regresjonsanalysen indikerer også at prognosen som settes har et selvstendig bidrag etter at GMA, gjennomsnittspoeng fra videregående skole og intervjuet er tatt høye for. Grunnen til at disse relativt sterke resultatene er funnet er, på samme måte som lederprognosen, vanskelig å dedusere seg frem til på grunn av den «tause kunnskapen» som benyttes av Sjøforsvaret på denne arenaen. En forklaring kan være at det også her benyttes grundig selekterte observatører, men hvem «den rette observatøren» er, er basert på meninger innad i gruppen (den samme tause kunnskapen), heller enn eksplisitte trekk som kan valideres. En slik strategi, selv om den i dag virker å fungere godt for denne gruppen, kan være sårbar. Hvis meninger i gruppen endres ved at nøkkelpersoner skiftes ut e.l., vet man ikke om dette vil ha positiv, negativ eller nøytral effekt på resultatene, nettopp fordi man ikke har analysert hva som er årsaken til den høye korrelasjonen. En slik analyse kan også resultere i at man finner uheldige grunner til de sterke resultatene, for eksempel ved at samme mennesker observerer på opptaket og er veiledere på utdanningen. Selv om dette eksempelet skal tilhøre sjeldenhetene, kan det, på grunn av mangel på klare retningslinjer knyttet til dette, ha blitt praktisert mer enn antatt. Konklusjonen på vurderingssenter blir derfor den samme som på intervjuet: det vil være nødvendig med en spesifikk analyse av seleksjonsmetoden før man kan konkludere med årsaken til de sterke resultatene.

Basert på de internasjonale retningslinjene for vurderingssentre (Rupp et al., 2015), har Forsvarets vurderingssenter et stort forbedringspotensial når det komme til strukturering og kvalitetssikring. Disse retningslinjene er satt både for å sikre etisk og rettferdig seleksjon, samtidig som validiteten skal bevares. Først og fremst anbefales det at hver simulering pre-testes og valideres, slik at overflødige simuleringer kan fjernes eller forbedres og slik at de beste simuleringene kan identifiseres. Skåren på simuleringene kan settes av en simuleringsansvarlig som observerer alle lagene på samme oppgave, noe som i større grad sikrer nivellering. Det anbefales også at de simuleringene som brukes er så jobbnære som mulig. Slik det gjøres i dag overlates mye til den individuelle observatør ved at en person følger det samme laget i en uke og bruker simuleringene til å få frem det de mener trengs for å kunne ta den endelige avgjørelsen. Dette er sårbart både på grunn av ulik praksis på tvers av observatører (dette til tross for at Sjøforsvaret bruker mye tid på nivelleringsdiskusjoner), og på grunn av at det krever svært dyktige observatører. En av de store fordelene med at observatøren i dag kan følge den samme gruppen i en uke er at de mest sannsynlig til få et bilde som ligner typisk prestasjon, heller enn optimal prestasjon. Dette fordi det er svært vanskelig å opprettholde et toppnivå over så lang tid. En mulig løsning for å utnytte styrkene av dagens metode og et standardisert vurderingssenter, er å gjøre den tre dager lange fellesdelen til vurderingssenteret og la det resterende være uendret. Dette vil kreve en ny validering og en analyse av hvordan lagets observatør og vurderingssenteret skal vektes.

Styrker, svakheter og feilkilder

Eksisterende internasjonal teori og empiri er tydelige på viktigheten av en grundig jobbanalyse når man skal selektere (Morgeson & Campion, 2000; Rumsey, 2012). Forsvaret har valgt å ikke gjøre slike jobbanalyser i dette tilfellet, men har samlet ekspertmeninger om hva god ledelse i Forsvaret er (en slags kompetanseanalyse). Den felles seleksjonen er fundert i at alle er tiltenkt en lederrolle, og disse kompetansene trengs hvis man skal lede, uansett

hvor i Forsvaret man befinner seg. En strategi mer i tråd med eksisterende teori vil være å analysere de ulike stillingene, deretter stille seg spørsmålet «Er det fellestrekk ved disse stillingene som gir grunnlag for en felles seleksjon?». En slik analyse vil også sikre at valideringen som gjøres i dag er basert på relevant kriteria, i tillegg til at den kan skape et utgangspunkt for utvikling av nye seleksjonsmetoder, slik som den amerikanske marinen gjorde på 90-tallet (Rumsey, 2012). Inspirasjon til jobbanalyse kan også hentes av nyere analyser gjort av offiserene i den amerikanske hæren (Paullin et al., 2014).

Denne studien tar også kort for seg problemet med indirekte diskriminering, men gjør ingen analyser på grad av dette i Sjøforsvarets seleksjon til GBK. Samtlige av metodene som brukes i seleksjonsprosessen i mer eller mindre grad forbundet med dette i internasjonal forskning. Det er derfor en stor sannsynlighet for at dette også er gjeldene for denne organisasjonen. Ettersom Forsvaret eksplisitt ønsker større mangfold i sin organisasjon, er dette er aspekt som er nyttig å være bevisst på når man velger sine metoder. Tilstedeværelse av indirekte diskriminering er også et argument for å bruke jobbanalyser, da dette vil tydeliggjøre om det er behov for å bruke en slik seleksjonsmetode. Dette gjør også jobbanalysen til et svært nyttig dokument hvis det skulle oppstå klagesaker hvor man må forsvare bruken av dagens metoder.

Den endelige karakteren til kriteriet «Skikkethet som militær leder» er en subjektiv vurdering gjort av nærmeste overordnede på slutten av den fire måneder lange utdanningen (supervisor rating). Slike kriterier er vist i tidligere studier å ha en noe lav inter-rater reliabilitet på .57 (Viswesvaran et al., 1996). Utdanningen og vurderingen i dette tilfellet er gjort på en mer strukturert metode enn hva man kan anta at eksisterer på den gjennomsnittlige arbeidsplass. I tillegg til dette er det en klar intensjon fra dag en om å sette en karakter på slutten av studiet, noe som medfører mer målrettet arbeid og nivellerende rådføring mellom

offiserene. I sum er dette grunner for å anta at reliabiliteten ikke er like lav i dette tilfellet, men det er vanskelig å anta dette før en analyse på dette blir gjennomført.

Begrensningen av variasjon i prediktorene er en feilkilde som vurderes å ha stor påvirkning i denne studien. Dette er i hovedsak begrunnet med det høye antallet prediktorer i ulike trinn fra sesjon del 1 til elevene fullfører Sjøforsvarets GBK. De ulike trinnene og metodene gjør at korrigering for dette vil være vanskelig å gjennomføre. I sum medfører dette at resultatene som rapporteres i denne studien er regnet som konservative.

Det er en høy grad av manglende data for flere av variablene, spesielt for gjennomsnittspoeng fra videregående skole og «skikkethet som militær leder. For gjennomsnittspoeng fra videregående skole er de manglende dataene spredd utover hele utvalget, mens for «skikkethet for militær leder» er det komplett data for noen kull og fullstendig fraværende for andre. Siden det ikke er funnet systematiske grunner for de manglende dataene som kan tenkes å forvri resultatene, er det regnet som et lite problem i dette tilfellet. Størst effekt vil det mest sannsynlig ha på den statistiske styrken til regresjonsanalysen med «skikkethet som militær leder» som kriterium. Denne analysen har det 91 personer i analysen. Selv om dette er lavt sammenlignet med resten av studien, er det fortsatt høyere enn anbefalte minstekrav (Stevens, 1996; Tabachnick & Fidell, 2013).

Det kan også være en metodisk svakhet å benytte seg av korrelasjonsstudier på variabler hvor et datapunkt kan være delvis avhengig av et annet (Stevens, 1996). I denne studien var både «skikkethet som militær leder» og offisersvurderingen fra vurderingssenteret observasjoner av individer i gruppe. Ettersom prestasjonen til enkeltpersonen kan være avhengig av dynamikken i gruppen, er det sannsynlig dette har hatt innvirkning på den endelige skåren. Dette er et aspekt som er vanskelig å unngå når man ønsker å se hvordan elevene leder en gruppe, men er forsøkt å ta høyde for ved å bevisstgjøre observatørene, i

tillegg til å anbefale forflytninger mellom gruppene hvis de opplever at dynamikken er spesielt uheldig for visse elever.

Det er også verdt å nevne noen av studiens styrker. Først og fremst er dette et omfattende studie som ser på samtlige kull over en periode på fem år. Muligheten til å kunne se på en så stor andel av elevene som har gjennomført GBK i Sjøforsvaret er med på å sikre representativiteten til funnene. Studien inkluderer også seleksjonsdata som strekker seg over lengre tid og som dekker mange relevante aspekt i seleksjon. I tillegg til dette er det funnet sterke resultater sammenlignet med eksisterende forskning, noe som øker sannsynligheten for at seleksjonen gjort på FOS er verdifull for GBK i Sjøforsvaret. Ettersom det tidligere ikke er gjort noen valideringsstudier på GBK i Forsvaret, bidrar denne studien med ny og nødvendig informasjon som kvalitetssikrer en seleksjonsprosess som hvert år påvirker svært mange karrierer. Studien bidrar også med konkrete implikasjoner for seleksjonsprosessen i fremtiden.

Implikasjoner

Det er blitt forsøkt å bruke eksisterende teori sammen med de konkrete funnene i denne studien, til å gi en konkret tilbakemelding som Sjøforsvaret kan ha praktisk nytte av. Den første og viktigste implikasjonen til studien er at Sjøforsvarets opptak til GBK fungerer svært godt, så det anbefales å i grove trekk fortsette med seleksjonsmetoden som er brukt i denne perioden. Et økonomisk argument for å fjerne eller blindt redusere denne seleksjonsmetodikken står derfor i fare for å virke mot sin hensikt, ved at mindre produktive elever blir selektert (Hull,1928; Hunter & Hunter, 1984; Judiesch & Schmidt, 2000; Terpstra & Rozell, 1993). Studien finner også et behov for å analysere Sjøforsvarets metoder grundigere, spesielt for lederprognosen i intervjuet og vurderingssenteret, ettersom det er uklart hvorfor metodene fungerer så bra som de gjør. En slik analyse kan avdekke at deler av metoden er overflødig og bør kuttes.

Denne studien har vist at skoleprognosen i intervjuet ikke bidro med selvstendig forklart varians. Det er anbefalt at intervjuet kun bør fokusere på lederprognosen, og heller la skoleprognosen være en matematisk formel som inkluderer GMA og gjennomsnittspoeng fra videregående skole. Denne endringen vil enten frigjøre ressurser ved å korte ned intervjuet, eller tid som kan brukes til å styrke lederprognosen ytterligere. Det kan være utfordrende å lage en formel som ivaretar de individuelle forskjellene hos elevene (for eksempel har noen videreutdanning og noen ikke), men så lenge formelen predikerer prestasjonen i større eller samme grad som den kliniske vurderingen, så kan det regnes som et verdifullt alternativ.

Videre impliserer denne studien at vektingen Sjøforsvaret i dag gjør av de forskjellige prediktorene, kan forbedres fra et validitetsperspektiv. I dag vektes vurderingssenteret 70 %, skoleprognosen 20 % og de fysiske testene 10 % i det endelige opptaket. Dette betyr at lederprognosen kun benyttes til negativ seleksjon (en test man kun må stå på, uten at skillet mellom de moderate, de gode og de beste benyttes). Denne studien anbefaler å la lederprognosen, sammen med vurderingssenteret og gjennomsnittspoeng fra videregående skole, vekte inn i en «skikkethet som militær leder» - prognose, og at GMA og gjennomsnittspoeng fra videregående skole vekter inn i en «hovedkarakter» - prognose. Hvordan disse to prognosene vektes mot hverandre er ikke et spørsmål denne studien kan besvare, men er mer et verdispørsmål som Sjøforsvaret må ta stilling til. I dag graderes hovedkarakteren og «skikkethet som militær leder» er bestått/ikke-bestått. Dette kan indikere at Sjøforsvaret vil tydeliggjøre de beste elevene, mens det holder å være en tilstrekkelig god leder. Hvis denne logikken skal videreføres til seleksjon, vil man sette et minstekrav på «skikkethet som militær leder» - prognosen og velge de beste innenfor «hovedkarakter» - prognosen som også oppnår dette minstekravet.

Fremtidig seleksjon innenfor det militære yrket

I 2014 ga den anerkjente journalen «Military Psychology» ut et spesialnummer som fokuserte på seleksjon i militære settinger. Denne utgivelsen la vekt på utviklingen de siste hundre årene, hvor feltet er i dag og hva fokuset kommer til å være de nærmeste årene. Blant artiklene i denne utgaven ble det argumentert for verdien av å fokusere på «seleksjon for å beholde», heller enn bare «seleksjon for å prestere» (White, Rumsey, Mullins, Nye, & LaPort, 2014). I tillegg til dette har militær forskning vist fremskritt innenfor ikke-kognitive seleksjon- og klassifiseringsmetoder som for eksempel personlighet (Stark et al., 2014) og interesser (Ingerick & Rumsey, 2014). For kognitive tester er det støtte for at mer spesifiserte testbatterier enn GMA kan være verdifullt for å øke treffsikkerheten (Rumsey & Arabian, 2014).

Konklusjon

Samlet sett viser studien at alle prediktorene på og før FOS korrelerer med en eller begge kriteriene ved endt utdanning. Når studien ser nærmere på bidraget til FOS, etter at de eksisterende testene er tatt høyde for (gjennomsnittspoeng fra videregående skole og GMA), kommer det frem at skoleprognosen på intervjuet ikke bidrar med selvstendig forklaringskraft og at FOS totalt sett ikke bidrar med noe forklart varians av hovedkarakter etter endt utdanning. Bildet er derimot svært annerledes når kriteriet er «skikkethet som militær leder». I dette tilfellet viser både lederprognosen fra intervjuet og vurderingssenteret moderat korrelasjon, noe som trosser de svake funnene gjort av tidligere studier på FOS (Kjenstadbakk, 2012; Isaksen, 2014). Det kommer også frem at vurderingssenteret bidrar med signifikant forklart varians selv etter at alle andre tester er tatt høyde for. Oppsummert er det i denne studien gjennomsnittspoeng fra videregående skole og GMA som predikerer hovedkarakter, og gjennomsnittspoeng fra videregående skole, skoleprognosen fra intervju og vurderingssenteret som predikerer «skikkethet som militær leder».

Referanser

Andersen, N. (2004). Editorial – the dark side of the moon: Applicant perspectives, negative psychological effects (NPEs), and candidate decision making in selection. *International Journal of Selection and Assessment*, *12*, 1-8. doi: 10.1111/j.0965-075X.2004.00259.x

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion- related validity of assessment center dimensions. *Personnel Psychology*, *56*, 125-154. doi: 10.1111/j.1744-6570.2003.tb00146.x

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*, 561-589. doi: 10.1111/j.1744-6570.1999.tb00172.x

Caretta, T. R., Teachout, M. S., Ree, M. J., Barto, E. L., King, R. E., & Michaels, C. F. (2014). Consistency of the relations of cognitive ability and personality traits to pilot training performance. *The International Journal of Aviation Psychology, 24*, 247-264. doi: 10.1080/10508414.2014.949200

Chamorro-Premuzic T., & Furnham, A. (2010). *The psychology of personnel selection*. Cambridge: Cambridge University Press.

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2. Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, M. (2009). *Personnel selection: Adding value through people* (5. Ed.). Chichester: Wiley-Blackwell.

Dakin, S. &, Armstrong, J. S. (1989). Predicting job performance: A comparison of expert opinions and research findings. *International Journal of Forecasting, 5*, 187-194. doi:10.1016/0169-2070(89)90086-1

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: an assessment of published and non-published correlation matrices. *Personnel Psychology, 65*, 221-249. doi: 10.1111/j.1744-6570.2012.01243.x

Dany, F., & Torchy, V. (1994). Recruitment and selection in Europe: policies, practices and methods. In C. Brewster, & A. Hegewisch (Eds.), *Policy and practice in european human resource management: The Price Waterhouse Cranfield Survey* (pp. 68-85). London: Routledge.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment centre: an efficient tool or a hammer to kill a fly? *Personnel Psychology*, *55*, 827-849. doi: 10.1111/j.1744-6570.2002.tb00131.x

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment centre ratings: a meta-analysis. *Journal of Applied Psychology*, *93*, 685-691. doi: 10.1037/0021-9010.93.3.685

Dobson, P., & Williams, A. (1989). The validation of the selection of male British Army officers. *Journal of Occupational Psychology, 62*, 313-325. doi: 10.1111/j.2044-8325.1989.tb00502.x

Driskell, J. E., & Olmstead, B. (1989). Psychology and the military: Research applications and trends. *American Psychologist, 44*, 43-54. http://dx.doi.org/10.1037/0003-066X.44.1.43

Duval, S. J. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.) *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester: Wiley.

Eid, J., Lescreve, F., & Larsson, G. (2012). An International Perspective on Military Psychology. In J. H. Laurence, & M. D. Matthews (Eds.), *The Oxford Handbook of Military Psychology* (pp. 114-128). Oxford University Press.

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*, 283-291.

Feltham, R. (1988). Assessment centre decision making: judgemental vs. mechanical. *Journal of Occupational Psychology*, *61*, 237-241. doi: 10.1111/j.2044-8325.1988.tb00287.x

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment centre validity. *Journal of Applied Psychology*, *72*, 493-511. doi: 10.1037/0021-9010.72.3.493

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography, *Intelligence*, *24*, 13-23. http://dx.doi.org/10.1016/S0160-2896(97)90011-8

Guilford, J. P. (1988). Some changes in the Structure-of-Intellect Model. *Educational & Psychological Measurement*, *48*, 1-4. doi: 10.1177/001316448804800102

Hausknecht, J. P., Day, D. V. &, Thomas, S. C. (2004). Applicant reactions to selection procedures: an updated model and meta-analysis. *Personnel Psychology*, *57*, 639-683. doi: 10.1111/j.1744-6570.2004.00003.x

Hardison, C.M. & Sackett, P.R. (2007). Kriterienbezogene Validitat des Assessment Centers: lebendig und wohlauf? In H. Schuler (Ed.) *Assessment Center zur Potenzialanalyse* (pp. 192-202). Hogrefe: Gottingen.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervision performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi: 10.1111/j.1468-2389.2007.00399.x

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184-190. http://dx.doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897-913.

http://dx.doi.org/10.1037/0021-9010.86.5.897

Hull, C. L. (1928). *Aptitude Testing*. New York: World book company. http://dx.doi.org/10.1037/11019-000

Hunter, J. E. (1980). *Test validation of 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: US Department of Labor.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*, 340-362. http://dx.doi.org/10.1016/0001-8791(86)90013-8

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

http://dx.doi.org/10.1037/0033-2909.96.1.72

Ingrerick, M., & Rumsey, M. (2014). Taking the measure of work interests: Past, present, and future. *Military Psychology, 26*, 165-181. http://dx.doi.org/10.1037/mil0000045

International Test Commission (2001). International Guidelines for Test Use, *International Journal of Testing*, 1(2), 93-114. doi: 10.1207/S15327574IJT0102_1

Isaksen, N. (2014). *Seleksjon til Luftforsvaret*. Manuskript under utarbeidelse.

Jones, A., & Harrison, E. (1982). Prediction of performance in initial officer training using reference reports. *Journal of Occupational Psychology*, *55*, 35-42. doi: 10.1111/j.2044-8325.1982.tb00076.x

Judge T. A., & Higgins, C.A. (1998). Affective disposition and the letter of reference. *Organizational Behavior and Human Decision Processes*, *75*(3), 207-221. doi:10.1006/obhd.1998.2789

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M.R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621-652. doi: 10.1111/j.1744-6570.1999.tb00174.x

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87,* 765-780

Judge, T. A., Colbert, A. E., & Ilies, R. (2004). Intelligence and leadership: A quantitative review and test of theoretical propositions. *Journal of Applied Psychology, 89*, 542-552

Judiesch, M. K., & Schmidt, F. L. (2000). Between-worker variability in output under piece-rate versus hourly pay systems. *Journal of Business Psychology*, *14*, 529-551. doi: 10.1023/A:1022932628185

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgement* (pp. 49-81). New York: Cambridge University Press.

Kjenstadbakk, T. J. (2012). Seleksjon til befalsskolen: En evaluering av seleksjonssystemets predikative validitet. Masteroppgave ved Stabbskolen. Oslo: Forsvarets Høgskole.

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C. (2006). Incremental validity of assessment centre ratings over cognitive ability tests: a study at the executive management level. *International Journal of Selection and Assessment*, *14*, 360-371. http://dx.doi.org/10.1111/j.1468-2389.2006.00357.x

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*, 1060-1072. doi: 10.1037/a0034156

Køber, P. K. (2015). Velger Forsvaret de rette ungdommene?: En analyse av seleksjon, gjennomføring og frafall i førstegangstjeneste 2010-2014. FFI 2014/02174. Hentet fra http://www.ffi.no/no/Rapporter/14-02174.pdf

Lai, L. (2004). *Strategisk kompetansestyring* (2. Ed.). Bergen: Fagbokforlaget

LeBreton, J. M., Scherer, K. T., & James, L.R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgement. *Industrial and Organizational Psychology, 7*, 478-500. doi: 10.1111/iops.12184

Lent, R. H., Aurbach, H. A., & Levin, L.S. (1971). Predictors, criteria, and significant results. *Personnel Psychology, 24*, 519-533. doi: 10.1111/j.1744-6570.1971.tb00374.x

Lievens, F. (2007). Research on selection in an international context: current status and future directions. In M. M. Harris (Ed.). *Handbook of research in International Human Resource Management* (pp. 107-123). Mahwah, NJ: Erlbaum.

Martinussen, M. (2005). Seleksjon av flygere og flygeledere. *Tidsskrift for Norsk Psykologforening, 42*, 291-300.

Martinussen, M. & Torjussen, T. M. (2004). Initial validation of a computer-based assessment battery for pilot selection in the Norwegian Air Force. *Human Factors and Aerospace Safety, 4*, 233- 244.

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105- 115. doi: 10.1024/1421-0185/a000012

Merton, R. K. (1948). The self-fulfilling prophecy. The Antioch Review, 8, 193-210. doi: 10.2307/4609267

McDaniel, M. A., Hartman, N. S., & Grubb, W. L. (2007). Situational judgement tests, response instructions, and validity: a meta-analysis. *Personnel Psychology*, *60*, 63-91. doi: 10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: a comprehensive review and meta-analysis. *Personnel Psychology*, *60*, 63-91. doi: 10.1037/0021-9010.79.4.599

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior, 21*, 819-827. doi: 10.1002/1099-1379(200011)21:7<819::AID-JOB29>3.0.CO;2-I

Murphy, K. R. (2002). Can conflicting perspectives on the role of g in personnel selection be resolved? *Human Performance*, *15*, 173-186. doi:10.1080/08959285.2002.9668090

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence of unconscious alteration of judgement. *Journal of Personality and Social Psychology, 35*, 250-256. http://dx.doi.org/10.1037/0022-3514.35.4.250

Paullin, C., Legree, P. J., Sinclair, A. L., Moriarty, K. O., Campbell, R. C., & Kilcullen, R. N. (2014). Delineating officer performance and its determinants. *Military Psychology, 26*, 259-277. http://dx.doi.org/10.1037/mil0000051

Ployart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: strategies for reducing racioethnic and sex subgroup difference and adverse impact in selection. *Personnel Psychology*, *61*, 153-172. doi: 10.1111/j.1744-6570.2008.00109.x

Ree, M. J., & Earles. J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, *1*, 86-89. doi: 10.1111/1467-8721.ep10768746

Reilly, R. R., & Chao, G. R. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*, 1-62. doi: 10.1111/j.1744-6570.1982.tb02184.x

Robertson, I. T. & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74, 441-472.* doi: 10.1348/096317901167479

Rom, E., & Kalderon, Y. (2013). The predictive role of simulations in assessing military performance. *Military Psychology, 25*, 402-411. doi: 10.1037/mil0000006

Roth, P. L., BeVier, C. A., Bobko, P., Switzer F. S., & Taylor. P. (2001b). Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. *Personnel Psychology, 54*, 297-330. doi: 10.1111/j.1744-6570.2001.tb00094.x

Roth, P. L., BeVier, C. A., Switzer III, F. S., & Schippmann, J.S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, *81*, 399-406. doi: 10.1037/0021-9010.81.5.548

Roth, P. L., & Bobko, P. (2000). Collage grade point average as a personnel selection device: ethnic group differences and potential adverse impact. *Journal of Applied Psychology, 85*, 399-406. doi: 10.1037/0021-9010.85.3.399

Roth, P. L., & Huffcutt, A. I. (2013). A meta-analysis of interviews and cognitive ability: Back to the future?. *Personnel Psychology, 12*, 157-169. http://dx.doi.org/10.1027/1866-5888/a000091

Rumsey, M. G. (2012). Military selection and classification in the United States. In J. H. Laurence, & M. D. Matthews (Eds.), *The Oxford Handbook of Military Psychology* (pp. 129-147). Oxford University Press.

Rumsey, M. G., & Arabian, J. M. (2014). Military Enlistment Selection and Classification: Moving Forward. *Military Psychology, 26*, 221-251. http://dx.doi.org/10.1037/mil0000040

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., …Jackson, D. J. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *17*, 243-253. doi: 0149206314567780

Rynes, S. L., Orlitzky, M. O., & Bretz Jr., R. D. (1997). Experienced hiring versus college recruiting: practices and emerging trends. *Personnel Psychology*, *50*, 309-339. doi: 10.1111/j.1744-6570.1997.tb00910.x

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. *Journal of Applied Psychology, 88*(6), 1068-1081. doi: 10.1037/0021-9010.88.6.1068

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: why there cannot be a debate. *Human Performance*, *15*, 187-210. doi: 10.1080/08959285.2002.9668091

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.

http://dx.doi.org/10.1037/0033-2909.124.2.262

Schmidt, F.L., & Hunter, J.E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975-2001. In Murphy, K.R. (Ed.), *Validity generalization: A critical review* (pp. 31-65). Mahwah: Lawrence Erlbaum.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*, 162-173. doi: 10.1037/0022-3514.86.1.162

Skoglund, T. H., Martinussen, M., & Lang-Ree, O. C. (2014). *Papir versus PC. Tidsskrift for Norsk Psykologforening, 51*, 450-452.

Skorstad, E. (2015). *Rett person på rett plass: Psykologiske metoder i rekruttering og lederutvikling* (2. Ed.). Oslo: Gyldendal Norsk Forlag AS.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Famer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26(3)*, 153-164. http://dx.doi.org/10.1037/mil0000044

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, *2*, 1-5. doi: 10.1111/1467-8721.ep10770441

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3. Ed.). Mahwah, NJ: Lawrence Erlbaum.

Stokke, M. (2000). *Bruk av tjenesteuttalelsen i seleksjon – problemer og muligheter*. Krigsskolen, Oslo: Hærens kompetansesenter for ledelse og utdanning. Hentet fra http://www.academia.edu/1295256/Bruk_av_tjenesteuttalelsen_i_seleksjon_-_problemer_og_muligheter

Sundet, J. M., Barlaug D. G., & Torjussen, T. M. (2004). The end of the Flynn effect?: A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32*, 349-362. doi:10.1016/j.intell.2004.06.004

Sundet, J. M., Tambs, K., Magnus, P., & Berg, K. (1988). On the question of secular trends in the heritability of IQ test scores: A study of Norwegian twins. *Intelligence*, *12*, 47-59. doi:10.1016/0160-2896(88)90022-0

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6.Ed.). Boston: Pearson Education.

Taylor, P. J., Pajo, K., Cheug, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel psychology*, *57*, 745-772. doi: 10.1111/j.1744-6570.2004.00006.x

Terpstra, D. E., Mohamed, A. A., & Kethley, R. B. (1999). An analysis of Federal court cases involving nine selection devices. *International Journal of Selection and Assessment*, *7*, 26-34. doi: 10.1111/1468-2389.00101

Terpstra, D. E., & Rozell, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology*, *46*, 27-48. doi: 10.1111/j.1744-6570.1993.tb00866.x

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: a meta-analytic review. *Personnel Psychology*, *44*, 703-742. doi: 10.1111/j.1744-6570.1991.tb00696.x

Thomassen, E. (2014). Evaluering av tjenesteuttalelsen i Forsvaret. En studie av dens egnethet som verktøy. Masteroppgave ved Stabsskolen. Oslo: Forsvarets Høgskole. http://hdl.handle.net/11250/216617

Thorndike, R. L. (1949). Personnel selection; test and measurement techniques. 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 Psykologiforening, 36*, 772-779.

Vik, J. S. (2013). Har seleksjon noen betydning? En studie av seleksjonens prediktive validitet. Masteroppgave ved Institutt for sosiologi, statsvitenskap og samfunnsplanlegging. Tromsø: Universitetet i Tromsø. http://hdl.handle.net/10037/5310

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574. doi:10.1037/0021-9010.81.5.557

White, L. A., Rumsey, M. G., Mullins, H. M., Nye, C. D., & LaPort, K. A. (2014). Toward a new screening paradigm: Latest army advances. *Military Psychology, 26*, 138-152. http://dx.doi.org/10.1037/mil0000047

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, *61*, 275-290. doi: 10.1111/j.20448325.1988.tb00467.x

Zimmerman, R. D., Triana, M. D. C., & Barrick, M. R. (2010). Predictive criterion-related validity of observer ratings of personality and job-related competencies using multiple raters and multiple performance criteria. *Human Performance*, *23*, 361-378. doi: 10.1080/08959285.2010.501049