Papir versus PC

Tom Hilding Skoglund , Monica Martinussen og Ole Christian Lang-Ree

Papir versus PC

Hvert år gjennomgår 20 000 personer intelligenstester på den militære sesjonen. Forsvaret benytter i økende grad PC-basert testing til fordel for papirog-blyant, men samsvarer resultatene fra de to formatene?

Forsvaret har en lang tradisjon med møysommelig klassifisering av militært personell ut fra psykologiske forhold (se f.eks. Hansen, 2006; Martinussen, 2005; Torjussen & Hansen, 1999). Kroneksempelet er testene på alminnelig evnenivå (AE) som siden tidlig på 1950-tallet har vært benyttet på sesjon, den primære seleksjonsarenaen i Forsvaret. Tidligere ble alle menn i hvert årskull testet, men dette forandret seg da den nye sesjonsordningen trådte i kraft fra 2009. En todelt sesjonsordning ble innført, hvor også jenter fikk sesjonsplikt. Basert på spørreskjemaresultater hvor hele årskull svarer (sesjon del 1), velges de antatt best egnede ut til et fysisk oppmøte på et sesjonssenter for videre seleksjon (sesjon del 2). Testene på AE gjennomføres på sesjonens andre del, hvor om lag 15 000 gutter og 5000 jenter møter opp hvert år. Gitt det store antallet som testes, har PC-basert testing vært et ønskelig alternativ til den tradisjonelle papir-og-blyant- metoden. I tillegg til å være åpenbart økonomibesparende, reduserer PC-basert testing også risikoen for menneskelig svikt i standardiseringsprosedyrer. Den teknologiske løsningen som muliggjorde PC-basert testing, kom samtidig med den nye sesjonsordningen, og enkelte sesjonssentre gikk da bort fra papir-og-blyant-metoden. Fra et testfaglig og testetisk ståsted bør det imidlertid undersøkes empirisk om to ulike testformater kan påvirke testtakerens prestasjoner før ordningen med PC-basert testing rulles ut over hele landet.

I tillegg til å være åpenbart økonomibesparende, reduserer PC-basert testing også risikoen for menneskelig svikt i standardiseringsprosedyrer

Formålet med denne studien er å sammenligne de to versjonene (papir- vs. PCbasert) av Forsvarets sesjonstester. Basert på funn fra et utvalg internasjonale studier forventer vi å finne relativt høyt samsvar mellom de to formatene.

Metode

Sesjonstestene består av tre deltester med tidsbegrensning: regneproblemer (U4), figurregler (U5) og ordlikhet (U6). U4 måler matematiske ferdigheter og resonneringsevne gjennom 30 praktiske regneoppgaver og varer i 25 minutter. U5 har en tidsbegrensning på 20 minutter, og består av 36

oppgaver som kartlegger abstrakt resonnering. Raven-testene og deltesten matriseresonnering i WAIS-III har likheter med U5. Endelig har U6 en varighet på 6 minutter, og måler ordforståelse gjennom 54 begreper. Alle tre deltester besvares ved hjelp av multiple-choice, og skåres i form av antall riktige svar. I praktisk bruk regnes dette om til Stanine-skårer (ni-delt skala), men for å oppnå større presisjon er råskårene anvendt i denne studien. Som et samlet mål på AE er gjennomsnittlig standardskår for de tre deltestene benyttet. En studie av Sundet og medarbeidere rapporterte en korrelasjon mellom sesjonstestene samlet sett og WAIS total IQ på 0,73, noe som støtter begrepsvaliditeten til sesjonstestene (Sundet, Tambs, Magnus & Berg, 1988). Test–retest-reliabiliteten ble undersøkt i samme studie, og man fant henholdsvis 0,84, 0,72 og 0,90 for de tre deltestene (Sundet et al., 1988).

Tabell 1 Gjennomsnittskårer og standardavvik for deltester (råskårer) og AE (gjennomsnittlig z-skåre), samt t-test og effektstørrelse (Cohens d) for forskjeller mellom testformatene (N = 439–440)

Tester	Papir		PC			
U4	15,64	4,83	15,71	4,66	–0,64	0,03
U5	26,28	6,05	26,69	4,80	–1,91	0,09
U6	29,30	8,37	29,03	8,32	1,27	0,06
AE	0,00	0,84	–0,02	0,84	0,09	0,00

I løpet av våren 2010 gjennomgikk 440 personer sesjonstestene to ganger med henholdsvis papir-og-blyant-versjon og PC-versjon. Av disse var 391 (89 %) menn og 49 (11 %) kvinner. Hver enkelt respondent svarte på de to versjonene samme dag med to til tre timers mellomrom. Rekkefølgen på formatene ble endret hver uke, mens datasamlingen pågikk over seks måneder slik at omtrent halvparten av utvalget tok papirversjonen først, mens den andre halvparten tok PC-versjonen først.

Tabell 2 Interkorrelasjoner (r) og intra-klasse-korrelasjoner (ICC) mellom papir-og-blyant-versjon og PC-versjon for alle deltester og AE (gjennomsnittlig z-skåre) (N = 439–440)

	r	*ICC*
U4	0,87**	0,87
U5	0,68**	0,68
U6	0,86**	0,86

AE	0,85**	0,85

Resultater og diskusjon

Analysene viste ikke signifikante gjennomsnittforskjeller på de to testformatene hverken på deltestnivå eller testene samlet, samt at effektstørrelsene i form av Cohens *d* er meget lave (se tabell 1). Videre viser korrelasjonene mellom de to testformatene (tabell 2) at skårene på U4 (regneproblemer) og U6 (ordlikhet) samvarierte høyt, mens skårene på U5 (figurregler) samvarierte noe mer moderat. Det er en høy korrelasjon mellom de to formatene på deltestene samlet (AE), r = 0,86. Intraklasse- korrelasjoner ICC (1,1) (Shrout & Fleiss, 1979) mellom de to testformatene ble også beregnet (presentert i tabell 2) og angir om både nivået og rekkefølgen på kandidatene samstemmer for de to formatene. Resultatene stemte i stor grad med de ordinære korrelasjonene mellom de to formatene.

Funnene våre er for det meste i samsvar med andre studier av ekvivalens mellom papirversjon og dataversjon på kognitive evnetester, men det er enkelte metodiske forskjeller. En undersøkelse av Raven Standard Progressive Matrices rapporterte også ikke-signifikante forskjeller i råskårgjennomsnittene på papir -og PC-versjon (Arce-Ferrer & Guzman, 2009). Forfatterne oppga ikke korrelasjoner mellom de to versjonene. Et dansk forskerteam rapporterte derimot en signifikant forskjell i gjennomsnittlige testskårer mellom papirversjon og PC-versjon på danskenes sesjonstest: «Børge Priens prøve» (Teasdale, Hartmann, Pedersen & Bertelsen, 2011). Imidlertid skyldtes nok dette metodiske forhold, spesielt fraværet av muligheten for en motbalanserings-design og et langt test–retest-intervall. Teasdale et al. (2011) testet om lag 100 offisersskole-kandidater med PC-versjonen av Børge Priens prøve, den samme testen som kandidatene hadde tatt ved sesjon noen år tidligere med papir og blyant. I gjennomsnitt gikk det 3 år mellom test og retest, og det viste seg, kanskje ikke overraskende, at det ble prestert signifikant bedre på den PC-baserte retesten. Korrelasjonen mellom test og retest ble oppgitt å være 0,77, noe som samsvarer godt med våre funn.

Resultatet vårt samsvarer også godt med resultatet fra en eldre metaanalyse av ekvivalensstudier med papir- og dataversjoner på kognitive tester (Mead & Drasgow, 1993). I 36 ulike undersøkelser med tidsbegrensede tester, hvorav 22 av studiene benyttet deltester fra det amerikanske Forsvaret (ASVAB – Armed Services Vocational Aptitude Battery), fant man en gjennomsnittlig korrelasjon på 0,72. Forfatterne hevdet at motoriske momenter hos testtakerne kan ha

sørget for at sammenhengen ikke var enda sterkere for tidsbegrensede tester, i og med at det krever en annen fysisk respons å registrere noe med tastatur/mus enn med en blyant. Motorikkhypotesen er interessant og intuitivt appellerende, men det kan problematiseres at flertallet av studiene som metaanalysen baserte seg på, ble utført på 1980-tallet. I tillegg til en omfattende utvikling av skjermkvalitet og brukervennlighet av tastatur og pekere er det også rimelig å anta at flertallet i befolkningen har blitt mer vant til å anvende dataverktøy.

Williams og McCord (2006) fant en korrelasjon på 0,59 på en ekvivalensstudie med Raven Standard Progressive Matrices, som er noe lavere enn våre funn. Til forskjell fra Raven-studien til Arce-Ferrer og Guzman (2009) hadde denne studien langt færre deltakere og et lengre test– restest-intervall på 53 dager. Forfatterne konkluderte med at papirversjon og PCversjon på Raven-testen i det store og hele gir samme resultat.

Studien vår har enkelte svakheter. Test og retest måtte av praktiske grunner foregå på samme dag, da sesjonens del to kun innebærer én dags oppmøte på sesjonssenteret. Dette er et betydelig kortere intervall enn det som er brukt i studiene til Williams og McCord (2006) og Arce-Ferrer og Guzman (2009), samt i intervallet på 3 uker, som ble brukt i reliabilitetsberegningene til WAIS-IV (Teasdale et al., 2011). Vi kan anta at læringseffekten er mer aktuell ved kortere intervaller, spesielt ettersom det kan være lett for testtakeren å huske de første oppgavene (som oftest er de letteste) når retesten starter. Slik kan testtakeren oppnå bedre tid til å løse de mer avanserte oppgavene, som gjerne kommer mot slutten av testene. Studien kunne videre vært tjent med å kartlegge rekkefølgeeffekter, men vi hadde ikke forberedt en presis registrering av formatrekkefølge når datainnsamlingen startet opp.[1] Motbalanseringen veier i noen grad opp for rekkefølgeeffekter. Styrker ved studien utover motbalansering er antallet forsøkssubjekter, samt inkluderingen av intra- klasse-korrelasjoner.

Konklusjon

Funnene viser at det ikke er vesentlige forskjeller i prestasjoner på papirversjonen og PC-versjonen av Forsvarets sesjonstester, og samsvarer dermed godt med internasjonale ekvivalensstudier av tester

[1] Forsvarets personellsystem har registrert klokkeslett for når PC-testene var ferdigstilt; vi forsøkte derfor å utlede formatrekkefølge ved å fordele utvalget under og over medianklokkeslettet. Imidlertid ble denne fremgangsmåten for usikker til at vi tok den med i videre analyser.

på kognitivt evnenivå. Testen U5 (figurregler) har noe lavere samvariasjon enn U4 (regneproblemer) og U6 (ordlikhet). Samvariasjonen skiller seg imidlertid ikke i nevneverdig grad fra de oppgitte internasjonale Raven-studiene. Forfatterne vurderer derfor at U5 sin samvariasjon er uproblematisk. Vi konkluderer derfor med at Forsvaret vil operere innenfor testfaglig og testetisk forsvarlige rammer når den økonomibesparende overgangen til PC-basert sesjonstesting lanseres over hele landet i årene fremover. Vi regner også med at overgangen resulterer i bedre standardiseringsprosedyrer, og slik sett bidrar til en ytterligere profesjonalisering av den militære sesjonen.

Referanser

Arce-Ferrer, A. J. & Guzman, E. M. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices Test. *Educational and Psychological Measurement, 69*, 855–867. doi: 10.1177/0013164409332219

Hansen, I. (2006). *Bidrag til psykologitjenestens historie i Forsvaret fra 1946–2006.* Militærpsykologiske meddelser, nr. 25. Oslo: Forsvarets skolesenter.

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

Mead, A. D & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458. doi: 10.1037//0033–2909.114.3.449

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86*, 420–428. doi: 10.1037//0033–2909.86.2.420

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.

Teasdale, T. W., Hartmann, P. V. W., Pedersen, C. H. & Bertelsen, M. (2011). The reliability and validity of the Danish draft board cognitive ability test: Børge Prien?s Prøve. *Scandinavian Journal of Psychology, 52*, 126–130. doi: 10.1111/j.1467–9450.2010.00862.x

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

Williams, J. E. & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior, 22*, 791–800. doi: 10.1016/j.chb.2004.03.005