

Testkvalitetsprosjektet – del 2: Tester med behov for kvalitetstiltak

Tanya Marie Ryder

Barne- og ungdomspsykiatrisk avdeling, Sykehuset i Vestfold Postboks 2168, 3103

Tønsberg, Norge

tanya.ryder@siv.no

Testkvalitetsprosjektet viser at mange hyppig brukte tester i Norge, har kvalitetsmangler knyttet til overholdelse av opphavsrett, oversettelsesmetodikk, norske normeringsdata, reliabilitet og validitet.

Norsk psykologforening (NPF) har lenge vært bekymret for tester og testbruk i Norge, basert på tidligere undersøkelser som har funnet store kvalitetsmangler (Selvik, 2007; Vaskinn et al., 2012; Vaskinn et al., 2010[IB2] ; Kornør et al., 2015). Høsten 2017 inngikk NPF et samarbeid med Helsedirektoratet og Folkehelseinstituttet (FHI) for å gjennomføre en oppfølgende kartlegging av norske psykologers holdninger til testkvalitet og av deres testbruk. Hensikten var å få en oppdatert oversikt over utfordringene i testfeltet og å identifisere hvilke tester som var mye brukt og samtidig hadde problemer knyttet til opphavsrett, oversettelse, norske normdata, reliabilitet eller validitet, for å kunne vurdere mulige korrigerende tiltak. I testkvalitetsprosjektets del 1 (Ryder, 2021) presenteres resultatene relatert til psykologers holdninger til testkvalitet, hvor hyppig ulike tester anvendes samt informasjon om utfordringer knyttet til ulike kvalitetstiltak fra helsemyndighetenes side. I denne artikkelen presenteres de 10 testene som vurderes å ha størst kvalitetsproblemer blant de 40 mest brukte testene i Norge.

Bakgrunn

International Test Commission (ITC) publiserte i 1996 for første gang retningslinjer med spesifiserte krav til testkvalitet, blant annet protokoll for oversettelser og språklig og kulturell tilpasning av tester (v. 2.4; ITC, 2017). Testutviklere skal i henhold til ITCs retningslinjer dokumentere adekvat tilpasning av tester med hensyn til språk og kultur, og de skal videre dokumentere statistiske analyser som i en tilfredsstillende grad understøtter den tilpassete versjonens psykometriske egenskaper. Det er viktig å være klar over at originalversjonens normer og dokumentasjon på reliabilitet og validitet ikke automatisk kan overføres til tilpassete versjoner (ITC, 2017, s. 15). Videre beskriver retningslinjene at oversettelse bør gjennomføres av minst to oversettere som har ekspertise på begge gjeldende språk og kulturer, på innholdet i testen og når det gjelder generelle prinsipper i testing. For visse tester vil det være påkrevd å engasjere en lingvist for å tilstrekkelig ivareta ulike språklige og kulturelle aspekter av oversettelsen. Det er verdt å merke seg at de fleste psykologiske tester som brukes i Norge, er opprinnelig utviklet i engelskspråklige land. Mange av disse har ikke blitt tilpasset norske forhold i samsvar med ITCs retningslinjer, noe som ytterligere svekker testkvaliteten og dermed påliteligheten av resultatene.

I tillegg til standarder for testkvalitet har ITC også retningslinjer for testbruk (v. 1.2; ITC, 2001) som ble vedtatt i 2000 som gjeldende for norske psykologer. Her heter det at testbrukere skal «velge teknisk godt utformede tester som passer til situasjonen» og «unngå bruk av tester som har inadekvat eller uklar støtte i teknisk dokumentasjon» (ITC, 2001, s. 13). Videre har testbrukere ansvar

for å respektere åndsverkloven (opphavsrett). For at norske psykologer skal kunne utøve forsvarlig testpraksis, er de avhengige av lett tilgjengelig kvalitetsinformasjon om tester brukt i Norge. Denne undersøkelsen vil kartlegge informasjonsgrunnlaget som norske testbrukere har tilgang til for å gjøre slike kvalitetsvurderinger.

Metode

Testundersøkelsen

Som en del av testkvalitetsprosjektet ble det samlet inn data om hvilke tester som brukes, fra en liste med over 350 tester av alle typer. Psykologene ble bedt om å indikere hvor ofte de brukte hver test: flere ganger i uken, flere ganger i måneden, flere ganger i kvartalet, sjeldnere enn hvert kvartal og sjeldnere enn hvert år. For mer detaljert informasjon om bakgrunnen for prosjektet, metodene for utarbeidelse av undersøkelsen og resultatene knyttet til psykologers testholdninger og testbruk, se Ryder (2021).

Informanter

Vi sendte undersøkelsen elektronisk til 4915 medlemmer av NPF i februar 2018. Studenter, pensjonister og psykologer som nylig hadde mottatt en annen undersøkelse fra NPF, ble ekskludert. Vi analyserte 1406 svar, noe som ga en svarprosent på 29.

Tabell 1. De 40 mest brukte testene

Test	Antall psykologer som bruker testen minst flere ganger i kvartalet (N = 1406)
1. Beck Depression Inventory-II (BDI-II)	299
2. Beck Anxiety Inventory (BAI)	223
3. Mini Internasjonalt Nevropsykiatrisk Intervju plus (M.I.N.I plus)	207
4. Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV)	198
5. Symptom Checklist 90-Revised (SCL-90-R)	197
6. Adult ADHD Self-Report Scale-v1.1 (ASRS-v1.1)	195
7. Mini Internasjonalt Nevropsykiatrisk Intervju (M.I.N.I.).	192
8. Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II)	159
9. Behavior Rating Inventory of Executive Function (BRIEF; parent and teacher version; 5-18 år)	147
10. Achenbach System of Empirically Based Assessment – Child Behavior Checklist (ASEBA – CBCL)	144
11. Montgomery Åsberg Depression Rating Scale (MADRS)	144
12. Alcohol Use Disorder Identification Test (AUDIT)	142
13. Achenbach System of Empirically Based Assessment – Teacher Report Form (ASEBA – TRF)	138
14. Conners Continuous Performance Test 3 (Conners CPT 3)	134
15. Kartlegging av traumeerfaringer-Barn/Foreldre (KATE-B/F)	130
16. Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)	119
17. Achenbach System of Empirically Based Assessment – Youth Self-Report (ASEBA – YSR)	118
18. Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V)	117
19. Diagnostisk intervju for utredning av ADHD hos voksne (DIVA 2.0)	117
20. Behavior Rating Inventory of Executive Function-Self-Report Version (BRIEF-SR; 11-18 år)	111
21. Drug Use Disorders Identification Test (DUDIT)	110
22. Achenbach System of Empirically Based Assessment – Caregiver-Teacher Report Form (ASEBA – C-TRF)	105
23. Delis-Kaplan Executive Function System (D-KEFS)	101
24. Behavior Rating Inventory of Executive Function-Adult Version (BRIEF-A; 18–90 år)	99
25. Child and Adolescent Trauma Screen (CATS)	97
26. ADHD Rating Scale-IV: Hjemmeversjon (ADHD-RS-IV: Hjemme)	95
27. ADHD Rating Scale-IV: Skoleversjon (ADHD-RS-IV: Skole)	91
28. Rey Complex Figure Test and Recognition Trial (RCFT)	88
29. California Verbal Learning Test-II (CVLT-II)	83
30. Schedule for Affective Disorders and Schizophrenia for School Aged Children (6–18 years): Present and Lifetime Version (Kiddie-SADS-PL)	78
31. Wisconsin Card Sorting Test (WCST)	70
32. Behavior Rating Inventory of Executive Function-Preschool Version (BRIEF-P; 2–5:11 år)	69
33. Vineland Adaptive Behavior Scales-Second Edition (Vineland-II)	68
34. Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV)	64
35. Norsk revidert Mini Mental Status Evaluering (MMSE-NR3)	61
36. Autism Diagnostic Interview-Revised (ADI-R)	61
37. Wechsler Memory Scale-III (WMS-III)	58
38. Impact of Event Scale-Revised (IES-R)	58
39. FAS Test (COWAT; verbal flyt)	57
40. Autism Spectrum Screening Questionnaire (ASSQ)	56

Bruksfrekvens ble målt ved å summere antall responser for tester som ble brukt minst «flere ganger i kvartalet», og alle tester ble rangert deretter. Analysen begynte med den mest brukte testen og ble deretter gjennomført sekvensielt nedover listen etter bruksfrekvens. Innenfor prosjektets ramme ble de første 40 testene (se Tabell 1) gjennomgått for å finne de 10 testene som ble vurdert til å ha forholdsvis størst kvalitetsproblemer.

Den relative alvorlighetsgraden av testers kvalitetsmangler ble vurdert opp mot andre tester i listen. Det er mange faktorer som inngår i en kvalitetsvurdering av en test, og hvilke som gjelder, vil avhenge av formålet med testen og formatet på den. Av den grunn er det vanskelig å gradere kvalitetssvakheter ved en test, da det er flere faktorer som kan virke sammen. Tester uten lovlig bruksrett ble imidlertid vurdert som ikke aktuelle å fortsette å anvende i Norge. Mangel på norske normer ble vurdert som et alvorlig kvalitetsproblem, da det vil være utfordrende å tolke resultatenes grad av avvik fra målpopulasjonen uten disse. Psykometrisk informasjon om reliabilitet og validitet ble vurdert som det nest viktigste aspektet å undersøke, forutsatt at norske normer foreligger, for å kunne vurdere om testen klarer å belyse de egenskapene man er interessert i å måle i den populasjonen, og at testen klarer å gjøre det på en konsekvent måte. Det er viktig å merke seg at reliabilitet og validitet vil være avhengig av den intenderte målgruppens normer. Manglende informasjon om hvordan tester ble oversatt, ble vurdert som det minst alvorlige kvalitetsproblemet i forhold til de andre, selv om dette klart kan ha betydning for validiteten av testresultatene.

Testinformasjon angående opphavsrett, oversettelse, norske normer, reliabilitet og validitet ble forsøkt innhentet fra kilder som vanlige testbrukere har tilgang til, som testmanualer, internettsidene til testleverandører og fagdatabaser som PsykTestBarn, FHI og Helsebiblioteket. I tillegg ble det søkt i artikkeldatabaser som PsycINFO, PubMed og Google Scholar ved bruk av søkeord relatert til testenes normer, reliabilitet og validitet samt oversettelsesprosesser. Noe informasjon fra norske studier ble vurdert som irrelevant, som studier med et lavt antall deltagere eller med lite generaliserbare kliniske populasjoner, og disse ble derfor ikke inkludert i analysen eller i testenes omtale under.

Resultater

Bakgrunnsvariabler

Av de rundt 51 % av respondentene som oppga demografisk informasjon, var 72 % kvinner og 54 % spesialister. Nesten alle respondenter oppga arbeidssted og fagfelt. De fleste jobbet i spesialisthelsetjenesten (62 %), derav 48 % i psykisk helsevern (29 % med voksne og 19 % med barn og unge).

10 tester med alvorlige kvalitetsproblemer og anbefalte tiltak

Listen under representerer 10 tester med vesentlige kvalitetsmangler som bør prioriteres for kvalitetstiltak. Det er viktig å merke seg at det finnes flere enkle kvalitetsmangler ved andre tester blant de 40 mest brukte testene. Det at en test ikke er omtalt i denne listen, betyr ikke at alle kvalitetsmål er innfridd. Prioritering av testene for kvalitetstiltak er basert på forfatterens helhetsvurdering av testfrekvens, egenskapene ved kvalitetsmanglene og praktisk betydning av manglene for testbruk. Rekkefølgen av testene her er etter bruksfrekvens.

1. Mini International Neuropsychiatric Interview (M.I.N.I.) og M.I.N.I. plus (Sheehan et al., 1998). M.I.N.I., og den utvidede utgaven M.I.N.I. plus, er strukturerte kliniske intervjuer som brukes i forbindelse med kartlegging av psykiatriske symptomer hos voksne. Intervjuene er basert på det amerikanske psykiatriske diagnosesystemet Diagnostic and Statistical Manual of Mental Disorders –

Fourth Edition (DSM-IV) / Fifth Edition (DSM-5) (American Psychiatric Association, 2000, 2013) med tilnærmede koder fra Den internasjonale statistiske klassifikasjonen av sykdommer og beslektede helseproblemer – 10. revisjon (ICD-10) (Verdens helseorganisasjon, 2020) som også brukes i USA. I intervjuets informasjon til norske brukere av M.I.N.I. versjon 6.0.0. påpekes det at grunnet enkelte forskjeller i diagnostiske kriterier mellom de to diagnostiske systemene, er kjennskap til ICD-10 diagnosekriterier påkrevet.

M.I.N.I. ble globalt oppdatert i 2016 for å samsvare med DSM-5 (APA, 2013), og det var Standard M.I.N.I. versjon 7.0.2 som var den siste godkjente norske oversettelsen ved undersøkelsestidspunkt (M.S. Edieux ved Mapi Research Trust, personlig kommunikasjon, 11. juli 2018). Bruken av M.I.N.I. krever en lisensavtale fra rettighetsinnehaveren Sheehan, og det er ikke tillatt å distribuere M.I.N.I. (Mapi Research Trust, 2021; Harm Research Institute, 2021; D.V. Sheehan, personlig kommunikasjon, 10. desember, 2020). Januar 2021 fant man publisert for åpen bruk på Helsebiblioteket.no utdaterte norske versjoner av M.I.N.I. basert på det foreldete diagnosesystemet DSM-IV-TR: M.I.N.I. versjon 6.0.0 fra 2009 og M.I.N.I. plus 5.0.0 fra 2005; samt M.I.N.I. screen 5.0.0 fra 2001 basert på DSM-IV (APA, 1994; Helsebiblioteket, 2021). Det er ikke beskrevet i testprotokollene hvordan M.I.N.I. 6.0.0 og M.I.N.I. plus 5.0.0 ble oversatt til norsk. Sheehan anbefaler ikke bruk av foreldete M.I.N.I.-utgaver, da de ikke er i tråd med den nyeste godkjente versjonen, som har gjennomgått store endringer (D.V. Sheehan, personlig kommunikasjon, 10. desember 2020).

Litteratursøket førte ikke frem til studier om reliabilitet eller validitet av den norske Standard M.I.N.I. 7.0.2. Basert på informasjon fra testens utvikler om at foreldete versjoner ikke skal brukes, samt Kunnskapssenterets tidligere vurdering av at disse versjonene manglet validitetsdata (Kunnskapssenteret, 2011a og 2011b i Egeland et al., 2015), bør Standard MINI 7.0.2 erstatte M.I.N.I. 6.0.0 i praksis. Videre bør M.I.N.I. tilleggsmoduler for DSM-V, som har mer detaljerte diagnostiske spørsmål for ulike lidelser, samt barne- og ungdomsversjonene (M.I.N.I. Kid-versjoner), gjøres tilgjengelig på norsk med tillatelse fra Sheehan. Til slutt bør reliabiliteten og validiteten av den gjeldende M.I.N.I. 7.0.2 undersøkes i en norsk populasjon, og det samme vil gjelde for M.I.N.I. tilleggsmoduler og M.I.N.I. Kid-versjoner i fremtiden.

2. Adult ADHD Self-Report Scale Symptom Checklist version 1.1 (ASRS) og ASRS screener (Kessler et al., 2005). Instrumentet ASRS omfatter et klinisk intervju for voksne med 18 symptomer av hyperkinetisk forstyrrelse basert på DSM-diagnosesystemet. ASRS screener er et kort selv-rapport skjema for voksne som inneholder de 6 symptomene som best predikerer hyperkinetisk forstyrrelse, og som vil kunne gi grunnlag for videre utredning. Både intervjuet og screeninginstrumentet ble oversatt med Verdens helseorganisasjons standard oversettelses- og tilbakeoversettelsesprotokoll, som er i samsvar med anbefalt praksis. Skjemaene kan brukes fritt så lenge det vises til WHO's rettigheter på skjemaene. En systematisk undersøkelse av ASRS ble publisert i PsykTestBarn (Kornør et al., 2011c). Man fant støtte for begrepsvaliditet, men ikke dokumentasjon for informasjon om reliabilitet. Det ble heller ikke funnet norske normdata, kun gjennomsnittsskårer for en norsk ikke-offisiell versjon av ASRS (Halleland et al., 2009 i Kornør et al., 2011c). Kornør og kollegaer konkluderte at det ikke fantes tilstrekkelig informasjon om reliabilitet og validitet, og at det måtte derfor utøves varsomhet ved bruk i klinisk praksis. I 2020 fant heller ikke vi norske normdata eller informasjon om testens reliabilitet eller validitet. Selv i Nasjonal faglig retningslinje for ADHD skrev Helsedirektoratet (2016) at en høy skår på ASRS kan forekomme ved ulike psykiske og somatiske tilstander, og de understreker viktigheten av å foreta en grundig og helhetlig vurdering. Testutvikleren har også poengtert at ASRS screeningsinstrument trenger valideringskartlegging for målpopulasjonen (Kessler, i.d.). Da det mangler norske normer og informasjon om tekniske testegenskaper, samt etter faglige anbefalinger

fra Kessler (i.d.) og Kornør et al. (2011c), bør man utarbeide norske normer for ASRS. Således bør fremtidige studier av reliabilitet og validitet av ASRS med et norsk normgrunnlag prioriteres.

3. Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II; First et al., 1997). SCID-II er et klinisk intervju brukt i forbindelse med kartlegging av symptomer av personlighetsforstyrrelser hos voksne, basert på det utdaterte DSM-IV diagnosesystemet. American Psychiatric Association eier oversettelesrettighetene til SCID-II (M.B. First, personlig kommunikasjon, 14. januar, 2021), og ved undersøkelsestidspunktet fantes det ingen autorisert norsk utgave (P. Telikicherla ved American Psychiatric Association Publishing, personlig kommunikasjon, 07. desember 2020). APA er tydelig på at SCID-II verken skal brukes eller distribueres i Norge. Situasjonen med utstrakt bruk av ikke-autorisert SCID-II har vært kjent siden Selviks testundersøkelse i 2006, og budskapet ble gjentatt senere på websiden til NPF (Egeland et al., 2015). Likevel var norske SCID-II-protokoller, i strid med åndsverkloven, gjort tilgjengelig mange steder på internett, inkludert på Helsebiblioteket.no, og dens bruk er anbefalt i nasjonale veiledere (Helsedirektoratet, 2012, 2013). Gyldendal Akademisk har nå blitt rettighetshaver for SCID i Norge. De utga i 2019 den første autoriserte norske versjonen av SCID: SCID-5-PD (SCID-5-PF på norsk) (First et al., 2019), som bygger på DSM-5 (APA, 2013). Mens det ikke er beskrevet hvilke metoder som ble brukt til å oversette SCID-II til norsk, er SCID-5-PD oversatt etter anbefalte metoder i overensstemmelse med ITCs retningslinjer. Når det gjelder psykometri, skrev APA i sin manual og på nettsiden i 2020 at det ikke forelå data om reliabilitet eller validitet for SCID-5-PD (American Psychiatric Association, 2020). Manualen for SCID-5-PF viser kun til utenlandske studier av SCID-II som viste store forskjeller i reliabilitets- og validitetsverdier i ulike undersøkelser og for ulike diagnostiske kategorier. Siden enhver bruk av SCID-II utgjør brudd på opphavsrett, skal denne erstattes med SCID-5-PF (P. Telikicherla ved American Psychiatric Association Publishing, personlig kommunikasjon, 07. desember 2020). Fordi det mangler opplysninger om verktøyets psykometriske egenskaper, bør det gjennomføres reliabilitets- og valideringsstudier for SCID-5-PF i en norsk populasjon.

4. Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000). BRIEF er et sett med spørreskjemaer som brukes til å vurdere eksekutiv fungering i hverdagen hos individer i forskjellige aldersgrupper: BRIEF-P (Preschool version, for rapportering på barn 2–5 år; Gioia, Espy & Isquith, 2003), BRIEF-SR (Self-report version, for selvrappotering hos unge 11–18 år; Guy, Isquith & Gioia, 2004), BRIEF foreldreversjon og BRIEF lærerversjon for rapportering på barn og unge mellom 5 og 18 år (Gioia, Isquith, Guy & Kenworthy, 2000) samt BRIEF-A (Adult version, for selvrappotering hos voksne; Roth, Isquith & Gioia, 2005). Skjemaene består av flere delskalaer som omhandler ulike aspekter av reguleringsfunksjoner. Hogrefe utgir testen i Norge med norske skjemaer og en amerikansk manual med amerikanske normer. Ifølge informasjon gitt fra Hogrefe ble oversettelsen utført i samsvar med anbefalt praksis i 2004, og oversettelsen ble revidert i 2007 (Hogrefe, personlig kommunikasjon, 28. mai 2020). Norsk forskning på et utvalg av 9–10 åringer gir generell støtte for bruk av amerikanske normer for foreldre- og lærerversjonene, selv om middelverdiene på enkelte delskalaer (Monitorering i foreldreversjonen og Fleksibilitet og Planlegging/organisering i lærerversjonen) lå lavere i det norske utvalget enn i det amerikanske utvalget (Fallmyr et al., 2011; Sørensen et al., 2014). Det kan bety at enkelte problemområder ikke fanges opp hos norske barn og unge ved bruk av de amerikanske normene (Køhn et al., 2020). Samlet sett konkluderes det i PsykTestBarn at den norske foreldreversjonen og lærerversjonen av BRIEF har tilfredsstillende psykometriske egenskaper (henholdsvis Sørensen et al., 2014, og Køhn et al., 2020). Det ble derimot ikke funnet relevante norske valideringsstudier om testegenskaper for BRIEF-SR eller BRIEF-A.

En norsk studie på BRIEF-P fylt ut av foreldre viste at selv om skjemaet klarte å skille kliniske fra ikke-kliniske grupper, ga anvendelse av amerikanske normer i kliniske grupper forholdsvis lave / ikke signifikante skårer på alle de 5 delskalaene (Skogan et al., 2015). Denne svekkede sensitiviteten kan innebære at man ikke oppdager noen typer problematisk atferd ved bruk av de amerikanske normene. Grunnet store individuelle forskjeller samt overlapp mellom ulike diagnostiske grupper advarte Skogan et al. (2015), i likhet med BRIEFs utviklere, mot å bruke skjemaet som det eneste grunnlaget for diagnose.

Etter at datainnsamlingen for vår studie ble avsluttet i 2018, kom det ut en ny versjon i Norge, BRIEF-2 (Gioia et al., 2015), som ble oversatt i samsvar med standard protokoll. Normene er imidlertid fortsatt amerikanske, oppdatert i 2013 og 2014 i forbindelse med revisjon av skjemaet. Litteratursøket ga ingen relevante norske studier om testegenskaper for BRIEF-2. Oppsummert mangler det norske normer, og det er heftet usikkerhet ved ulike kvalitetsaspekter ved noen av BRIEF-skjemaene. Siden det har kommet en revidert utgave av testen, BRIEF-2, bør den versjonen av testen prioriteres for utarbeidelse av norske normer for alle aldersgrupper, samt studier av reliabilitet og validitet basert på et norsk normgrunnlag.

5. ADHD-Rating Scale-IV: Home and school versions (ADHD-RS-IV; DuPaul et al., 1998).

ADHD-RS-IV er et sett med skjemaer som fylles ut av enten foreldre (hjemmeversjon) eller lærer (skoleversjon) for å kartlegge de 18 symptomene av hyperkinetisk forstyrrelse i forbindelse med utredning og med utprøving av medisin for hyperkinetisk forstyrrelse. Skjemaene er basert på den utdaterte diagnosemanualen DSM-IV. Testen kan anvendes hos barn og unge i alderen 5 til 18 år. Persentilskårer utarbeides for hyperaktivitet/impulsivitet, uoppmerksomhet og totalskalaen, men disse er basert på representative amerikanske normer. Den norske versjonen er tilgjengelig i boken AD/HD: Et verktøy for kartlegging av barn og ungdom (Kvilhaug et al., 1998), der det står skrevet at tillatelse for oversettelse og bruk i Norge er gitt av R. Barkley i 1996. Kjøpere av boken har tillatelse til å kopiere skjemaet kun for personlig bruk. Imidlertid er oversettelsesmetoden ikke tilstrekkelig beskrevet for å kunne vurdere om den var i tråd med ITCs retningslinjer.

Skjemaene er i den originale engelskspråklige versjonen anerkjent som nyttige verktøy, og de er i utstrakt bruk. Men hjemme- og skoleversjoner ble vurdert i PsykTestBarn, og konklusjonen var at det verken finnes norske normer eller grunnlag for å kunne vurdere testens psykometriske egenskaper (Kornør et al., 2011a og 2011b). I 2016 kom ADHD-RS-V (DuPaul et al., 2016) ut i USA, oppdatert for DSM-V med hjemmeversjoner og skoleversjoner både for barn i alderen 5 til 10 år og for ungdommer i alderen 11 til 17 år. Litteratursøket ga ingen norske oversettelser eller norske studier av disse nye skjemaene. Grunnet mangel på norske normer samt publikasjon av den oppdaterte ADHD-RS-V bør de nyeste versjonene prioriteres for oversettelse og utarbeidelse av norske normer. Fremtidige studier av reliabilitet og validitet av ADHD-RS-V basert på norske normer bør prioriteres.

6. California Verbal Learning Test-II, Second Edition (CVLT-II; Delis et al., 2000).

CVLT-II er en test som måler verbal-læring og hukommelse. Testadministrator leser høyt ordlister som testpersonen skal repetere og gjenkjenne over ulike tidsintervaller. CVLT-II kan anvendes hos personer fra 16 år og oppover. Pearson utga i 2004 en norsk oversettelse av CVLT-II med et norsk manualsupplement (Lundervold et al., 2004). Mens manualen beskriver at ordlistene ble tilpasset norske forhold og kontrollert av en lingvist, er det for lite informasjon til å vurdere om ITCs retningslinjer for oversettelse er blitt tilstrekkelig fulgt. De kjønnsdelte normene er hentet fra USA. Norske studier med både ikke-kliniske deltagere og blandete kliniske grupper har ikke vist de kjønnsforskjellene man finner i det amerikanske normgrunnlaget, slik at anvendelse av de amerikanske normene hos norske kvinner kan gi misvisende resultater

for noen aspekter av hukommelse (Egeland et al., 2005; Langlo et al., 2015; Kanestrøm, 2017). Manualen inneholder ikke informasjon vedrørende reliabilitet eller validitet for den norske versjonen. Kunnskapssenterets systematiske litteratursøk av den norske CVLT-II fant noe støtte for begrepsvaliditet og konvergentvaliditet, men de fant ikke norske studier som ga tilstrekkelig informasjon om reliabilitet (Sjiveland et al., 2014). Mens testen har kommet ut i en ny versjon i USA, CVLT-3, har testleverandøren formidlet at de ikke har planer om å oppdatere den i Skandinavia (Pearson Clinical, personlig kommunikasjon, 17. juni 2020). Dette betyr at CVLT-II vil være den tilgjengelige versjonen i Norge de kommende årene. Grunnet manglende norske normer og usikkerhet rundt de psykometriske egenskapene til CVLT-II bør det utarbeides norske normer. Fremtidige studier av reliabilitet og validitet av CVLT-II basert på norske normer bør også prioriteres.

7. Wisconsin Card Sorting Test (WCST; Heaton et al., 1993). WCST måler mental fleksibilitet og evner til å bruke kognitive strategier i problemløsning ved at testadministrator gir tilbakemelding om organisering av ulike typer stimuli etter ulike regler. Testen kan administreres både i papirform og på datamaskin, og den kan brukes hos de som er minst 6 år gamle. Hogrefe utgir WCST med amerikanske normer og et norsk manualsupplement som inkluderer informasjon på norsk om administrasjon, skåring og tolkning samt preliminare normdata basert på 60 friske svensker mellom 20 og 55 år (Grant et al., 2004). Det finnes en del norske studier der WCST brukes i testing av kliniske grupper og kontroller, både av barn og voksne, men et litteratursøk ga ikke tilstrekkelig informasjon om psykometriske egenskaper. I en eldre valideringsstudie (Rodríguez-Aranda et al., 2006b) ble WCST administrert til 101 friske nordmenn mellom 20 og 88 år. Noen av resultatene ser ut til å være korrelert med alder og utdanningsnivå, og mangel på norske normer gjør tolkningen av resultatene usikker. Grunnet mangel på norske normer og tvil knyttet til testens psykometriske egenskaper anbefales det at man utarbeider norske normer for WCST samt prioriterer fremtidige studier av testens reliabilitet og validitet basert på norske normer.

8. Autism Diagnostic Interview-Revised (ADI-R; Rutter et al., 2003). ADI-R er et semi-strukturert klinisk intervju som er kjent som «gullstandarden» i kartlegging av symptomer innenfor autismspekterforstyrrelse (ASF). Intervjuet gjennomføres med foresatte eller omsorgspersoner for å innhente informasjon om atferd og utviklingshistorie hos de med en mental alder på minst 2 år. Testen er basert på kriterier i ICD-10 og i det utdaterte DSM-IV. ADI-R intervjuene tar i bruk beslutningsalgoritmer der grenseverdier indikerer mulig ASF. Det er nå også utviklet nyere amerikanske beslutningsalgoritmer for småbarn (Kim et al., 2012a). Hogrefe utgir ADI-R med norske protokoller oversatt etter anbefalte prosedyrer, men både manualen og normene med grenseverdier for beslutningsalgoritmene er amerikanske.

Halvorsen og kollegaers (2017) systematiske undersøkelse av ADI-R i PsykTestBarn fant ikke norske reliabilitetsstudier av ADI-R. Derimot fant de lavere sensitivitet for den norske ADI-R sammenlignet med amerikanske studier, og at testens validitet kan påvirkes av barnets alder, IQ og atferdsproblemer samt nivå av foreldrebekymring. Eksempelvis refererte de til Havdahl og kollegaers norske studie (2017) der bruk av de nye amerikanske algoritmene (Kim et al., 2012a) hos norske småbarn hvor foreldre rapporterte at de ikke var særlig bekymret for ASF, reduserte testens sensitivitet i en slik grad at over 40 % av småbarn med faktisk ASF ikke ble fanget opp. Mens disse nye algoritmene ikke utgis av testleverandøren i Norge, viser studien at testens validitet er spesifikk for den intenderte populasjonen.

Begrepsvaliditet, undersøkt ved den norske ADI-Rs grad av samsvar med det semi-strukturerte observasjonsverktøyet Autism Diagnostic Observation Schedule-Second Edition (ADOS-2; Lord et al., 2012), var tilfredsstillende (Havdahl et al., 2017). Halvorsen et al. (2017) mente at det var støtte

for bruk av de ovennevnte diagnostiske beslutningsalgoritmene for småbarn så lenge ADOS-2 også brukes i utredningen. I forbindelse med diagnostisk utredning anbefalte flere også at ADI-R brukes sammen med ADOS for å øke validiteten av testresultatene (Zander et al., 2015; Havdahl et al., 2017; Kim et al., 2012b). Zander et al. (2015) konkluderte at brukt isolert var ADOS-2 egentlig mer nyttig i forbindelse med diagnostikk enn ADI-R alene. I PsykTestBarn ble det oppsummert at det var behov for flere norske studier for å belyse testens reliabilitet og validitet, spesielt med tanke på de diagnostiske beslutningsalgoritmene (Halvorsen et al., 2017). Basert på mangel på norske normer og den generelle usikkerheten knyttet til testens psykometri anbefales det at det utarbeides norske normer og grenseverdier, og at fremtidige studier av testens reliabilitet og validitet prioriteres.

9. Wechsler Memory Scale-III (WMS-III; Wechsler, 2008). WMS-III brukes til utredning av ulike aspekter av læring, hukommelse og arbeidsminne hos personer som er 16 år og eldre. Pearson utgir testen med et norsk manualsupplement, men normene er amerikanske, datert tilbake til 1997. Det finnes også en WAIS-III/WMS-III teknisk manual på engelsk (Tulsky et al., 1997) for sammenligning av WMS-III-skårer med WAIS-III-skårer for å vurdere hukommelse i lys av intelligensnivå. Det beskrives at testen ble oversatt parallelt til norsk og svensk ved konsensus i arbeidsgruppen, men det er for lite informasjon til å vurdere om oversettelsesprotokollen er blitt fulgt, spesielt med tanke på bruk av en lingvist for å ivareta språktekniske aspekter ved denne hukommelsestesten.

Manualen beskriver den norske/svenske utprøvingen (N = 180) på utvalgte deltester fra WMS-III fra 2007, der 89 norske deltagere inngikk kun i utvalgte aldersgrupper: 25–29 år (n = 29), 45–54 år (n = 30) og 65–69 år (n = 30). Konklusjonen om at de amerikanske normene kunne brukes selv om nordmenn/svensker presterte noe høyere enn amerikanere på alle deltester unntatt to, Ansikter I og II og Familiebilder I og II, kan dermed være vanskelig for testbrukere å feste tillit til. Utprøvingen ble ikke gjennomført i alle aldersgruppene, og utvalget består av et lavt antall deltagere.

En studie er gjennomført med eldre, friske nordmenn fra 55 års alder (Bosnes et al., 2012). Forfatterne konkluderte at de amerikanske normene kunne brukes for den aldersgruppen, men at disse kan overestimere nordmenns generelle/utsatte hukommelse og dermed ikke fange opp mildere problemer med hukommelse. I en senere artikkel på de samme utvalgsdataene som over (Bosnes et al., 2016) viste resultatene at de kulturpregete deltestene, Ansikter I og Familiebilder I, ga avvik hos nordmenn på rundt $\frac{3}{4}$ standardavvik (SA). Man bør merke seg at den oppdaterte testversjon, WMS-IV (Wechsler, 2009), ikke lenger inneholder disse to deltestene. Nevropsykologiske tester brukes ofte til å måle fungering over tid, og derfor er det nødvendig at test-retest-reliabilitet oppgis for slike tester. Den amerikanske tekniske manualen for WMS-III viser at gjennomsnittlige indeksskårer øker mellom 0,33 og 1,0 SA ved retest etter 2–12 uker (Johnstone, 2010). Litteratursøket ga ingen studier som belyser grad av læringseffekt ved retesting hos nordmenn.

WMS har gjennomgått omfattende endringer i den fjerde amerikanske versjonen, og ovenstående usikkerhetsmomenter ved testens egenskaper viser behovet for en oppdatert norsk versjon. Testleverandøren besvarte ikke henvendelser om det er planlagt oppdateringer av WMS. Grunnet mangel på norske normer og behov for psykometrisk kvalitetsinformasjon anbefales det at man forsøker å få WMS-IV oversatt og tilpasset norske forhold, inkludert utarbeidelse av norske normer. Så bør fremtidige studier av testens reliabilitet og validitet basert på et norsk normgrunnlag prioriteres.

10. FAS Test (Controlled Oral Word Association Test; Benton et al., 1989 / Delis-Kaplan Executive Function System (D-KEFS), Verbal Flyt – Ordflyt; Delis et al., 2005). FAS Test inngår i D-KEFS, en sammensetning av flere velkjente nevropsykologiske testmetoder som brukes til å måle ulike aspekter av eksekutiv fungering hos barn og voksne. D-KEFS ble utviklet i USA for å standardisere

administrasjonsteknikker for flere tester samt å lage ett felles normsett for tolkning av resultatene. Alle tester i D-KEFS kan anvendes hos barn fra 8 års alder med unntak av «Ordtak», som kan anvendes fra 16 års alder. Pearson utgir testen med norske protokoller og et norsk manualsupplement (Delis et al., 2005), men det er viktig å merke seg at normene er amerikanske. I den norske manualen beskrives det at grunnet ulik utforming av hver av testene kunne noen av disse tas i bruk med kun oversettelse av instruksene, mens andre deltester krevde mer omfattende oversettelse og tilpasning til norsk språk og kultur, noe som ser ut til å være ivaretatt for det meste basert på begrenset informasjon i manualen.

FAS Test, eller Controller Oral Word Association Test, er deltesten som brukes til å måle fonetisk ordflyt, der testpersonen skal produsere flest mulig ord muntlig som oppfyller spesifikke kriterier innenfor en begrenset tidsramme, vanligvis innen 1 minutt. Teststimuli består i all hovedsak av enkle instruksjoner som testadministratoren leser for testpersonen, og derfor kreves det ikke en omfattende oversettelsesprosedyre slik som ved andre tester. Derimot er det vesentlig hvilke fonetiske kriterier som brukes i ulike språkpopulasjoner grunnet forskjellig forekomst av ord i ulike språk. Derfor spesifiserer Tombaugh et al. (1999) at normsettet kun må anvendes for det samme målspåket (for D-KEFS betyr det engelsk). En metaanalyse av 134 amerikanske og kanadiske studier om fonetisk ordflyt viser at ulike kriterier brukt selv innenfor det samme språket gir ulike vanskelighetsgrader (Barry et al., 2011). Det spesifiseres i D-KEFS-manualen at muligheten for å bruke sammensatte ord på norsk kan gi en forhøyet skår på FAS Test hos nordmenn. Studier viser at flere andre faktorer kan påvirke prestasjon på FAS Test, og dermed validitet av resultatene. Dette kan være ung alder (Delis et al., 2001), alder generelt (Rodríguez-Aranda et al., 2006a; Barry, 2011), utdanningsnivå (Loonstra et al., 2001; Tombaugh et al., 1999; Barry et al., 2011), lesehastighet og ordforståelse (Rodríguez-Aranda, 2003).

Den norske D-KEFS-manualen (Delis et al., 2005) beskriver at dataanalyser av voksne svenske og norske pasienter samt 57 friske voksne nordmenn, matchet med pasientgruppen med hensyn til utdanningsnivå, kjønnsfordeling og alder, støttet foreløpig bruk av de amerikanske normene. Imidlertid ble det funnet at middelværdien for den norske kontrollgruppen ($n = 57$) tilnærmet seg et SA over gjennomsnittet for den amerikanske normalpopulasjonen, og en gruppe norske pasienter med bipolar lidelse ($n = 14$) presterte mer enn 1 SA over den samme amerikanske normalpopulasjonen. Dessuten ble det ikke innhentet data for barn eller ungdom. Flere andre studier som anvendte FAS Test med friske norske barn og voksne, viser høyere gjennomsnittsverdier enn forventet sammenlignet med de amerikanske normene – i hvert fall for noen aldersgrupper (se Egeland et al., 2006, Rodríguez-Aranda et al., 2006; Andersson et al., 2010; Løvstad et al., 2012). Slike resultater fra bruk av utenlandske normer kan øke risikoen for at testpersoner med mildere svikt ikke oppdages. Kalechstein et al. (1998; i Delis, 2001) fant at flere tester av eksekutiv fungering som finnes i D-KEFS, har tidligere vist variasjon i standardiserte skårer seg imellom, og de poengterte at det derfor er nødvendig med et stort, enhetlig og representativt normsett for disse. Søket ga for øvrig ikke informasjon om testens reliabilitet for nordmenn. Grunnet mangel på norske normer og usikkerhet knyttet til testens psykometri anbefales det at man utarbeider norske normer for ikke bare FAS Test, men for hele D-KEFS, slik at forfatterens målsetting om ett samlet normsett for flere tester av eksekutiv fungering også kan gjelde i Norge. Så bør fremtidige studier av testens reliabilitet og validitet basert på norske normer prioriteres.

Diskusjon

Dette prosjektet viser at tydelig kvalitetsinformasjon om tester ofte er vanskelig å finne, og at det er vesentlige kvalitetsmangler med tester som brukes hyppig i Norge. Det omfatter bruksrett, oversettelsesmetodikk, norsk normeringsdata, reliabilitet og validitet. Av de ulike kvalitetskriteriene som ble undersøkt, var informasjon om hvem som var forfatteren av testen, enklest å fastslå.

Opplysninger om hvem som hadde oversatt en test, var ofte tilgjengelig, men informasjon om hvorvidt tillatelse til oversettelse foreligger, eller om faglig anbefalte prosedyrer for oversettelsesprotokoll (ITC, 2017) har blitt fulgt, manglet ved en del tester. Det var i noen tilfeller utfordrende å finne tydelig informasjon om hvorvidt det foreligger tillatelse til bruk eller distribusjon av tester i Norge. Slike opplysninger var heller ikke alltid oppdaterte etter som nyere testversjoner har blitt utviklet for å erstatte eldre versjoner. Vi avdekket også motstridende eller usikker informasjon om rettsbruk fra ulike kilder. I noen tilfeller tok vi kontakt med utenlandske testeiere for å skaffe eller bekrefte slik informasjon, men man bør ikke forvente at norske psykologer skal måtte gå til slike skritt for å være ansvarlige testbrukere.

Kvaliteten på informasjonen om norske normer var heller ikke alltid enkelt å tyde. Noen testforlag publiserte kun utenlandske normer, av og til med «preliminære» data på mindre norske eller svenske utvalg på visse aldersgrupper som skulle støtte bruk av utenlandske normer, men oppfølgingsstudier for å bekrefte konklusjonene manglet. Studier utført i kliniske miljøer og forskningsmiljøer der testresultater fra norske utvalg ble publisert, var som oftest basert på små kliniske grupper og ikke større representative referansegrupper som vil kunne gjøre data gjeldende som et normsett. Det er mange ulemper med å lene seg til data fra slike studier, da de kan gi misvisende resultater (Delis, 2001). Små utvalg gir usikker representativitet for den intenderte målgruppen. Færre deltagere i utvalgene kan også bety at disse samles i få, brede aldersgrupper der resultatene vil kunne være mindre sensitiv til alderseffekter ved visse egenskaper. Videre risikerer man at deltagere fra barne- og ungdomspopulasjonen utelates fra studier til fordel for voksne deltagere, og disse normene for voksne vil ikke kunne anvendes hos barn/unge.

Gjennom litteratursøket i undersøkelsen ble det funnet at få norske studier belyste ulike typer reliabilitet og validitet for tester. Psykometriske egenskaper for tester vil være unik for hver språkversjon, og derfor vil man ikke kunne støtte seg til originalversjonens rapporterte reliabilitet og validitet (ITC, 2017). Mens validiteten til et testverktøy som regel vil måtte vurderes skjønnsmessig av testbrukeren basert på formål med testen i ulike sammenhenger, er manglende informasjon imidlertid et stort problem.

Dette testkvalitetsprosjektet bekrefter tidligere funn om et misforhold mellom internasjonale testretningslinjer og statusen på testfeltet i Norge, og det er spesielt bekymringsfullt at noen av disse problemene har vært kjent over lang tid. Testforlag og andre testutgivere gjør tilgjengelig utenlandske testverktøy uten nødvendige tilpasninger for bruk i Norge, og det kan øke faren for feil i testresultatene. Alle 10 testene som er belyst i artikkelen, hadde problemer knyttet til norsk normering, reliabilitet og validitet. Det er ressurskrevende å utarbeide norske normer for tester og å gjennomføre studier som viser statistiske analyser som støtter den tilpassete versjonens psykometriske egenskaper, men det er avgjørende at dette arbeidet prioriteres for at man skal kunne ha tillit til testresultatene. Valideringsstudier som det er refererte til i denne artikkelen, har vist at bruk av utenlandske normer som er basert på populasjoner i land med betydelige forskjeller i demografi, språk og kultur sammenlignet med norske forhold, kan redusere validiteten til konklusjonene som utledes av testresultatene. Ofte slo dette ut i lavere sensitivitet, der spesifikke vansker som norske barn og voksne

faktisk har, ikke ble oppdaget, noe som kan resultere i alvorlige konsekvenser for testpersonene. Salg av tester med utenlandske normer og norske protokoller kan gi inntrykk av et kvalitetsstempel for bruk i Norge. Denne undersøkelsen viser at dette ikke alltid stemmer. Likevel bekrefter psykologer at mange ulike tester som mangler tilstrekkelig kvalitetsdokumentasjon, er i hyppig bruk i Norge.

Metodikken i dette prosjektet ansees å speile hvordan norske testbrukere ville søke informasjon om testkvalitet: via testmanualer, internettsidene til testleverandører, norske fagdatabaser og internasjonale artikkeldatabaser. Den er således ikke like systematisert og grundig som for eksempel undersøkelsene av måleegenskaper utført av PsykTestBarn/FHI, og noe relevant informasjon kan ha blitt oversett. I denne undersøkelsen kom det frem informasjon om flere ulike kvalitetsproblemer blant de 40 mest brukte testene, og andre fagpersoner vil kunne ha gjort ulike valg, etter en helhetsvurdering, når det gjelder hvilke tester man bør prioritere å forbedre. Likevel er det tydelig at det er stort rom for forbedring ved flere tester, og denne listen ansees som et godt utgangspunkt for det videre arbeidet.

Konklusjon

Denne undersøkelsen viser at det er behov for økt bevissthet rundt det ansvaret alle som er involvert med tester og testing, har. Informasjon om testkvalitet må både gjøres lettere tilgjengelig for testbrukere og den må kvalitetssikres og holdes oppdatert – noe som viser seg å være en formidabel, men viktig oppgave. Et større samarbeid mellom testleverandører, helsemyndigheter med ansvar for pasientsikkerhet og relevante fagmiljøer vil kunne stimulere til forbedringsinitiativer og etterfølgelse av gjeldende testretningslinjer. NPE, Helsedirektoratet og FHI har allerede vist interesse for å bidra. En nasjonal satsing på oversettelse, kulturell tilpasning, norsk normering og forskning om testegenskaper må til for å løfte testkvaliteten og øke tilliten til konklusjonene som baseres på bruk av psykologiske tester i Norge.

Merknad

Forfatteren har avgitt signert interessekonflikterklæring og oppgir ingen interessekonflikter. Innhold i artiklene er utarbeidet av forfatteren uten påvirkning fra de økonomiske bidragsyterne.

Takk

Takk til Helsedirektoratet som bidro med midler som gjorde prosjektet mulig. Det takkes for skrivestipend fra Barne- og ungdomspsykiatrisk avdeling ved Sykehuset i Vestfold. Takk til Norsk psykologforening: TPU-medlemmer (2017–2019) for innspill til undersøkelsen og analysemetoder, Andreas Høstmælingen og Ole Tunold for bistand med forberedelse og utsendelse av undersøkelsen samt datainnsamling og databearbeidelse, sentralstyret for skrivestipend og Mikael J. Sømhovd for kommentarer til manuskriptet.

Referanser

- American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders* (4. utg.).
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5. utg.). <https://doi.org/10.1176/appi.books.9780890425596>
- American Psychiatric Association Publishing. (2020, v1.0.0.19). *The Structured Clinical Interview for DSM-5®: Additional Support and Materials*. <https://www.appi.org/products/structured-clinical-interview-for-dsm-5-scid-5>

- Andersson, S., Lovdahl, H. & Malt, U. F. (2010). Neuropsychological function in unmedicated recurrent brief depression. *Journal of Affective Disorders*, 125, 155–164. <https://doi.org/10.1016/j.jad.2009.12.023>
- Barry, D., Bates, M. E. & Labouvie, E. (2008). FAS and CFL forms of verbal fluency differ in difficulty: a meta-analytic study. *Applied neuropsychology*, 15(2), 97–106. <https://doi.org/10.1080/09084280802083863>
- Benton, A. L. & Hamsher, K. (1989). *Multilingual Aphasia Examination* (2. utg.). AJA Associates.
- Bosnes, O. & Troland, K. (2012). Wechsler Memory Scale-III og Wechsler Adult Intelligence Scale-III utprøvd i et utvalg av HUNT 3-populasjonen. *Tidsskrift for Norsk psykologforening*, 49(5), 462–467.
- Bosnes, O., Troland, K. & Torsheim, T. (2016). A Confirmatory Factor Analytic Study of the Wechsler Memory Scale-III in an Elderly Norwegian Sample. *Archives of Clinical Neuropsychology*, 31, 12–17. <https://doi.org/10.1093/arclin/acv060>
- Delis, D. C., Kramer, J. H., Kaplan, E. & Ober, B. A. (2000). *Manual for the California Verbal Learning Test, Second Edition (CVLT-II)*. The Psychological Corporation.
- Delis, D. C., Kaplan, E. & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System*. Pearson Assessment. <https://doi.org/10.1037/t15082-000>
- Delis, D. C., Kaplan, E. & Kramer, J. H. (2005). *Delis-Kaplan Executive Function System, norsk manual supplement*. (T. Foss overs.). Pearson Assessment.
- DuPaul, G. J., Power, T. J., Anastopoulos, A. D. & Reid, R. (1998). *ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretation*. Guilford. <https://doi.org/10.1037/t00680-000>
- DuPaul, G. J., Power, T.J., Anastopoulos, A.D. & Reid, R. (2106). *ADHD Rating Scale-5 for Children and Adolescents: Checklists, Norms, and Clinical Interpretation*. Guilford Press.
- Egeland, J., Landrø, N. I., Tjemsland, E. & Walbækken, K. (2006). Norwegian Norms and Factor-Structure of Phonemic and Semantic Word List Generation. *The Clinical Neuropsychologist*, 20, 716–728. <https://doi.org/10.1080/13854040500351008>
- Egeland, J., Sundet, K., Landrø, N. I., Rund, B. R., Asbjørnsen, A. & Hugdahl, K. (2005). Validering av normer for oversatte tester av oppmerksomhet og hukommelse i et norsk normalutvalg. *Tidsskrift for Norsk Psykologforening*, 42, 99–105.
- Egeland, J., van Delft, C., Føllestad, H., Holen, A., Helland, S. & Engen Nilsen, D. Ø. (2015). *Det må være et myndighetsansvar å sette nødvendige prosesser i gang for å sikre helsevesenet tilgang på kvalitetssikrede verktøy*. <https://www.psykologforeningen.no/medlem/testbruk/uautoriserte-tester-og-kartleggingsverktoy>
- Fallmyr, Ø. & Egeland, J. (2011). Psykometriske egenskaper for den norske versjonen av Behavior Inventory of Executive Function (BRIEF). *Tidsskrift for Norsk psykologforening*, 48(4), 339–343.
- First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W. & Benjamin, L. S. (1997). *Structured clinical interview for DSM-IV axis II personality disorders, (SCID-II)*. American Psychiatric Association.
- First, M. B., Williams, J. B. W., Karg, R. S. & Spitzer, R. L. (2019). *Strukturert klinisk intervju for personlighetsforstyrrelser – DSM-5 (SCID-5-PF)*. (E. Haukeland overs.). Gyldendal Akademisk (original publisert 2016).
- Gioia, G. A., Espy, K. A. & Isquith, P. K. (2003). *BRIEF-P: Behavior Rating Inventory of Executive Function-Preschool Version*. Psychological Assessment Resources.
- Gioia, G. A., Isquith, P. K., Guy, S. C. & Kenworthy, L. (2000). *Behavior Rating Inventory of Executive Function*. Psychological Assessment Resources. <https://doi.org/10.1076/chin.6.3.235.3152>

- Gioia, G. A., Isquith, P. K., Guy, S. C. & Kenworthy, L. (2015). *BRIEF-2: Behavior Rating Inventory of Executive Function-Second Edition*. PAR Inc.
- Grant, D. A. & Berg, E. A. (2004). *Wisconsin Card Sorting Test. Manual Supplement, norsk versjon*. Hogrefe Psykologiförlaget AB.
- Guy, S., Isquith, P. & Gioia, G. (2004). *BRIEF-SR: Behavior Rating Inventory of Executive Function-Self Report Version*. Psychological Assessment Resources.
- Halvorsen, M. & Helverschou, S. E. (2017). Måleegenskaper ved den norske versjonen av Autism Diagnostic Interview-Revised (ADI-R). *PsykTestBarn*, 1(5).
- Harm Research Institute (16. januar, 2021). License Agreements for Use of Any MINI. <https://harmresearch.org///index.php/mini-international-neuropsychiatric-interview-mini/#License%20Agreements%20for%20Use%20of%20Any%20MINI>
- Havdahl, K. A., Bishop, S. L., Suren, P., Oyen, A. S., Lord, C., Pickles, A., von Tetzchner, S., Schjolberg, S., Gunnes, N., Hornig, M., Lipkin, W. I., Susser, E., Bresnahan, M., Magnus, P., Stenberg, N., Reichborn-Kjennerud, T. & Stoltenberg, C. (2017). The influence of parental concern on the utility of autism diagnostic instruments. *Autism Research* 10(10), 1672–1686. <https://doi.org/10.1002/aur.1817>
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G. & Curtiss, G. (1993). *Wisconsin card sorting test manual: Revised and expanded*. Psychological Assessment Resources.
- Helsedirektoratet. (2012). *Nasjonalt faglig retningslinje for utredning, behandling og oppfølging av personer med samtidig ruslidelse og psykisk lidelse – ROP-lidelser*.
- Helsedirektoratet. (2013). *Nasjonalt faglig retningslinje for utredning, behandling og oppfølging av personer med psykoselidelser*.
- Helsedirektoratet. (2016). *Nasjonalt faglig retningslinje for ADHD*.
- International Test Commission (2001). International Guidelines for Test Use [ITC retningslinjer for testbruk]. *International Journal of Testing*, 1(2), 93–114. (Norsk Psykologforening, oversetter, 2013). <https://www.intestcom.org/page/17> https://doi.org/10.1207/S15327574IJT0102_1
- International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. <https://www.intestcom.org/page/16>
- Johnstone, E. C., Owens, D. C., Lawrie, S. M., McIntosh, A. M. & Sharpe, M. (red.). (2010). *Companion to Psychiatric Studies* (8. utg.). Elsevier Ltd.
- Kanestrøm, H. (2017). California Verbal Learning Test (CVLT-II) og Brief Visuospatial Memory Test – Revised (BVMT-R): Undersøkelse av prestasjoner og samsvar i et klinisk utvalg. *Tidsskrift for Norsk Nevropsykologisk Forening*, 19(1), 10–16.
- Kessler, R. C. (i.d., hentet 15.04.2020). Adult ADHD Self-Report Scale (ASRS) Version 1.1: Background Information. https://www.hcp.med.harvard.edu/ncs/ftplib/adhd/background_memo_rev.pdf
- Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B. & Walters, E. E. (2005). The World Health Organization Adult ADHD Self-Report Scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, 35, 245–256. <https://doi.org/10.1017/S0033291704002892>
- Kim, S. H. & Lord, C. (2012a). New autism diagnostic interview-revised algorithms for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Autism and Developmental Disorders*, 42, 82–93. <https://doi.org/10.1007/s10803-011-1213-1>
- Kim, S. H. & Lord, C. (2012b). Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months

- of age. *Journal of Child Psychology and Psychiatry*, 53(2). <https://doi.org/10.1111/j.1469-7610.2011.02458.x>
- Kornør, H. & Bøe, T. (2011a). Måleegenskaper ved den norske versjonen av ADHD-RS-IV – ADHD Rating scale IV, Hjemmeversjon (ADHD-RS-IV Hjemme). *PsykTestBarn*, 1(8).
- Kornør, H. & Bøe, T. (2011b). Måleegenskaper ved den norske versjonen av ADHD Rating Scale IV, Skoleversjon (ADHD-RS-IV Skole). *PsykTestBarn*, 1(9).
- Kornør, H. & Hysing, M. (2011c). Måleegenskaper ved den norske versjonen av Adult ADHD Self Report Scale, 1.1 (ASRS). *PsykTestBarn*, 1(6).
- Kornør, H., Jozefiak, T., Hanssen-Bauer, K. & Hysing, M. (2015). Forsvarlig testbruk. *Tidsskrift for Norsk Psykologforening*, 52(7), 600–601.
- Kvilhaug, G., Høigaard, B., Rønnevold, T., Aase, H., Eilertsen, O., Rydin, S. A., Iglum, L., Farstad, A. L. & Johansen, E. B. (1998). *AD/HD: Et verktøy for kartlegging av barn og ungdom*. Novus forlag.
- Køhn, K. & Halvorsen, M. (2020). Måleegenskaper ved den norske lærerversjonen av Behavior Rating Inventory of Executive Function (BRIEF). *PsykTestBarn*, 10(5).
- Langlo, K. S. & Erdal-Aase, R. (2015). Testing av tegnspråklige døve med California Verbal Learning Test-II. *Tidsskrift for Norsk Psykologforening*, 10, 863–871.
- Loonstra, A. S., Tarlow, A. R. & Sellers, A. H. (2001). COWAT Metanorms Across Age, Education and Gender. *Applied Neuropsychology* 8(3), 161–166. https://doi.org/10.1207/S15324826AN0803_5
- Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K. & Bishop, S. (2012). *Autism diagnostic observation schedule–2nd edition (ADOS-2)*. Western Psychological Corporation.
- Lundervold, A. J. & Sundet, K. S. (2004). *Norsk Versjon. Manuals supplement. CVLT-II*. Psykologiförlaget AB.
- Løvstad, M., Funderud, I., Endestad, T., Due-Tønnessen, P., Meling, T. R., Lindgren, M., Knight, R. T. & Solbakk, A. K. (2012). Executive functions after orbital or lateral prefrontal lesions: Neuropsychological profiles and self-reported executive functions in everyday living. *Brain Injury*, 26(13–14), 1586–1598. <https://doi.org/10.3109/02699052.2012.698787>
- Mapi Research Trust. (16. januar, 2021). Mini-International Neuropsychiatric Interview (MINI). <https://eprovide.mapi-trust.org/instruments/mini-international-neuropsychiatric-interview#languages>
- Rodríguez-Aranda, C. (2003). Reduced Writing and Reading Speed and Age-related Changes in Verbal Fluency Tasks. *The Clinical Neuropsychologist* 17(2), 203–215. <https://doi.org/10.1076/clin.17.2.203.16508>
- Rodríguez-Aranda, C. & Martinussen, M. (2006a). Age-Related Differences in Performance of Phonemic Verbal Fluency Measured by Controlled Oral Word Association Task (COWAT): A Meta-Analytic Study. *Developmental Neuropsychology*, 30(2), 697–717. https://doi.org/10.1207/s15326942dn3002_3
- Rodríguez-Aranda, C. & Sundet, K. (2006b). The frontal hypothesis of cognitive aging: factor structure and age effects on four frontal tests among healthy individuals. *Journal of Genetic Psychology*, 167(3), 269–287. <https://doi.org/10.3200/GNTP.167.3.269-287>
- Roth, R. M., Isquith, P. K. & Gioia, G. A. (2005). *BRIEF-A: Behavior Rating Inventory of Executive Function-Adult version*. Psychological Assessment Resources.
- Rutter, M., Le Couteur, A. & Lord, C. (2003). *Autism diagnostic interview-revised*. Western Psychological Services.

- Ryder, T. M. (2021). Testkvalitetsprosjektet – del 1: Norske psykologers testholdninger og testbruk. *Tidsskrift for Norsk Psykologforening*, 58(1), 28–37.
- Selvik, A. (2007). *Forsvarlig bruk av tester, strukturerte intervjuer og spørreskjemaer – kvalitet og bruksrett i helsevesenet*. Internrapport Psykologforeningen.
https://ressursside.no/handbok-filer/forsvarlig_bruk_av_tester_npf.pdf
- Sheehan, D. V., Lecrubier, Y., Harnett-Sheehan, K., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G. (1998). The Mini International Neuropsychiatric Interview (M.I.N.I.): The Development and Validation of a Structured Diagnostic Psychiatric Interview. *Journal of Clinical Psychiatry*, 59(20), 22–33. <https://doi.org/10.1037/t18597-000>
- Siqveland, J., Sundseth, Ø., Dalsbø, T. K., Harboe, I. & Leiknes, K. A. (2014). *Måleegenskaper ved den norske versjonen av California Verbal Learning Test II (CVLT-II)* (Rapport fra Kunnskapscenteret nr. 7). Nasjonalt kunnskapssenter for helsetjenesten.
- Skogan, A. H., Zeiner, P., Egeland, J., Urnes, A. G., Reichborn-Kjennerud, T. & Aase, H. (2015). Parent ratings of executive function in young preschool children with symptoms of attention-deficit/-hyperactivity disorder. *Behavior and Brain Functions*, 11(16). <https://doi.org/10.1186/s12993-015-0060-1>
- Sørensen, L. & Hysing, M. (2014). Måleegenskaper ved den norske versjonen av Behavior Rating Inventory of Executive Function (BRIEF). *PsykTestBarn*, 2(6).
- Tombaugh, T. N., Kozak, J. & Rees, L. (1999). Normative Data Stratified by Age and Education for Two Measures of Verbal Fluency: FAS and Animal Naming. *Archives of Clinical Neuropsychology*, 14(2), 167–177. [https://doi.org/10.1016/S0887-6177\(97\)00095-4](https://doi.org/10.1016/S0887-6177(97)00095-4)
- Tulsky, D., Zhu, J. & Ledbetter, M. (1997). *WAIS-III/WMS-III Technical Manual*. Psychological Corporation.
- Vaskinn, A. & Egeland, J. (2012). Testbrukerundersøkelsen: En oversikt over tester brukt av norske psykologer. *Tidsskrift for Norsk psykologforening*, 49(7), 658–665.
- Verdens helseorganisasjon. (2020). *Den internasjonale statiske klassifikasjonen av sykdommer og beslektede sykdommer* (10. utg.). Direktoratet for e-helse. <https://ehelse.no/kodeverk/kodeverket-icd-10-og-icd-11>
- Wechsler, D. (2008). *Wechsler Memory Scale – Third edition, norsk versjon*. Pearson Assessment.
- Wechsler, D. (2009). *Wechsler Memory Scale-Fourth Edition*. Pearson.
- Zander, E., Sturm, H. & Bölte, S. (2015). The added value of the combined use of the Autism Diagnostic Interview–Revised and the Autism Diagnostic Observation Schedule: Diagnostic validity in a clinical Swedish sample of toddlers and young preschoolers. *Autism*, 19(2), 187–199. <https://doi.org/10.1177/1362361313516199>