

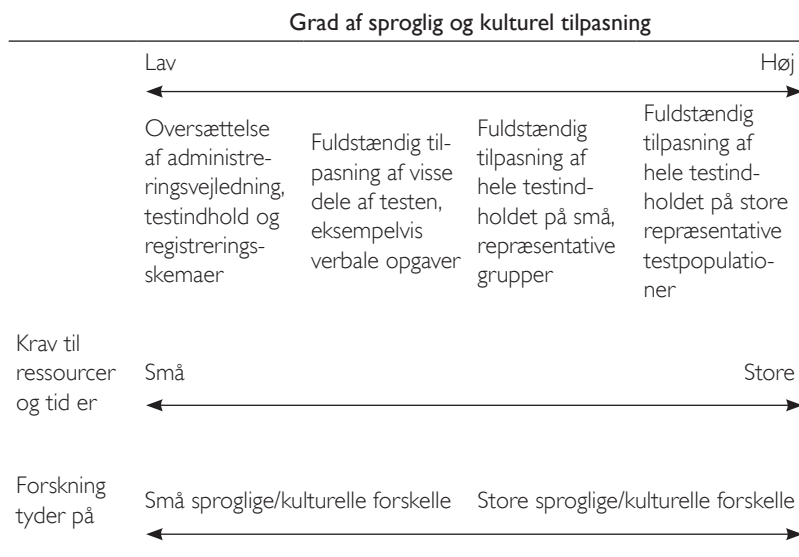
Afprøvning af den danske D-KEFS

Særlige overvejelser

At oversætte, tilpasse, teste og normere en test, der er udarbejdet på et andet sprog og i en anden kultur, indebærer i princippet, at testen konstrueres og udarbejdes på ny. Dette gælder især test, hvor man antager, at sprog og kulturelle fænomener er af stor betydning.

Formålet med en sproglig og kulturel tilpasning af testindhold er at sikre, at test fungerer tilfredsstillende i forbindelse med undersøgelse og diagnosticering. Dette indebærer, at testen skal have de tilsvarende psykometriske egenskaber med hensyn til reliabilitet og validitet som originalversionen, og at det teoretiske grundlag, som originalversionen er baseret på, er intakt i de nye sprogversioner.

Graden af tilpasning, som er nødvendig for en test, varierer og er afhængig af mange faktorer som eksempelvis dens karakter og udformningen af de enkelte opgaver, tid, ressourcer, samt erfaringer fra forskning og klinisk praksis. Et udviklingsarbejde kan indebære alt lige fra en fuldstændig tilpasning med normer baseret på store, statistisk repræsentative populationer og undersøgelser i kliniske grupper, til oversættelse og kulturel tilpasning af administreringsvejledningen, verbale opgaver samt mulighed for at registrere testpersonernes præstation (det vil sige oversættelse af registreringsskemaer og andet forbrugsmateriale), mens såvel belæg for reliabilitet og validitet samt normer i det tilfælde kan være hentet fra originalversionen af testen.



Figur 11.1 Faktorer, der hver især eller tilsammen påvirker den valgte grad af tilpasning.

Antallet af psykologer i klinisk praksis inden for det afgrænsede sprogområde, som Danmark udgør, influerer sammen med økonomiske og praktiske begrænsninger på overvejelserne om, hvor omfattende et afprøvnings- eller standardiseringsarbejde en given test kan bære.

Oversættelse og kulturel tilpasning stiller høje krav. Som hovedregel er selv en god oversættelse ikke tilstrækkeligt, da visse opgaver ikke uden videre kan overføres fra ét sprog til et andet og samtidig bevare det psykologiske konstrukt, som opgaven måler, og bevare den enkelte opgaves sværhedsgrad. Den vigtigste del af testudviklingsprocessen er derfor – med ovenstående begrænsninger in mente – at dokumentere testens psykometriske egenskaber: Reliabilitet, validitet og normernes gyldighed.

Ved dataindsamlinger, hvor formålet er at indsamle normer er udfordringen at sikre, at data indsamles blandt personer, som er repræsentative for befolkningen, som testen er beregnet til. Typisk vil man derfor kontrollere, om fordelingen af demografiske variable såsom alder, køn og uddannelsesniveau svarer til fordelingerne i den pågældende befolkningsgruppe. Tidligere erfaringer viser, at det er sværere at rekruttere yngre personer

til dataindsamlinger, og at det er sværere at rekruttere personer med kortere uddannelser end personer med længere uddannelser.

Et af udgangspunkterne for udviklingsarbejdet med den danske version af D-KEFS var at indsamle en begrænset mængde data for at undersøge, i hvilken udstrækning de amerikanske normer kan bruges blandt danskere. Antagelsen var, at der mellem de amerikanske og danske testpersoner vil være visse forskelle i præstation, og to grundlæggende spørgsmål blev derfor formuleret:

- Forekommer der forskelle i testpræstation, der kan tolkes som sande forskelle mellem de to populationer (USA og Danmark)?
- Har disse forskelle klinisk betydning?

Flere af D-KEFS-testene har tydelig verbal karakter, og præstationen i disse test beror i en vis grad på testpersonens sproglige formåen. En god oversættelse og kulturel tilpasning af opgaver i sådanne test er en forudsætning for anvendeligheden. Der findes dog også test i D-KEFS, hvor testindholdet hovedsagligt udgøres af figurer (inklusive tal og bogstaver), former og farver. Disse test hviler på samme teoretiske grundlag som andre ikke-verbale test (se beskrivelser af de enkelte test og deres historiske udvikling i kapitel 1).

Ved oversættelsen og tilpasningen af den danske D-KEFS er der taget to overordnede hensyn. Dels er den visuelle opbygning af stimulusark og stimulusbøger søgt tilnærmet de eksisterende svenske og norske versioner så meget som muligt af hensyn til anvendeligheden af dette materiale ved valideringsundersøgelsen i Danmark, og dels er de sprogligt ladede dele af testene søgt tilnærmet det amerikanske forlæg så meget som muligt.

Således er det semantiske indhold i stimulusord og begreber søgt bevaret bedst muligt, så eksisterende amerikanske scorings-eksempler og scoringsvejledninger kunne anvendes som direkte forlæg ved udarbejdelsen af de danske modstykker. Dette netop med det formål at kunne prøve de amerikanske normer på de danske testdata.

De psykometriske analyser har haft til formål at sammenligne resultater fra den amerikanske normering af D-KEFS med resultaterne fra den danske afprøvning. I hovedtræk blev

det undersøgt, i hvilken udstrækning korrelationsniveauer og korrelationsmønstre var forskellige mellem de to grupper (den amerikanske normgruppe og den danske valideringsgruppe), samt om gennemsnit og standardafvigelse var forskellige mellem de to grupper.

En overordnet hypotese blev formuleret:

Præstationen i den danske D-KEFS valideringsgruppe – målt ved gennemsnitlig præstation og spredning – adskiller sig ikke signifikant fra præstationen i den amerikanske normgruppe.

Denne overordnede hypotese har styret valget af analyser, således at potentielle forskelle i præstationer mellem den amerikanske og den danske gruppe er blevet fundet og vurderet i forhold til de to indledende spørgsmål – om forskelle findes, og om de i givet fald har klinisk betydning.

De psykometriske analyser fulgte denne plan:

- Hypoteseafprøvning og afklaring af det grundlæggende spørgsmål om gyldigheden af amerikanske normer i Danmark.
- Sammenligning af den interne struktur i de amerikanske og danske D-KEFS data.

Det er generelt accepteret blandt statistikere, at der er alvorlige begrænsninger forbundet med at anvende nulhypotese-test ved tolkning af testresultater (se fx Balluerka, Gómez, & Hidalgo, 2005; Cohen, 1990, 1994; Hubbard & Lindsay, 2008; Hunter, 1997; Kirk, 1996, 2001; Thompson, 1998). Chansen for at finde signifikante forskelle øges, jo større grupper man sammenligner, hvilket medfører, at selv små forskelle kan fremstå som signifikante. Dette har resulteret i en opfordring til at rapportere konfidensintervaller og effektstørrelsesmål i stedet for kun statistiske signifikans og p -værdier (American Educational Research Association, 2006; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). På baggrund af dette præsenteres derfor konfidensintervaller og effektstørrelser samt i visse tilfælde værdier for statistisk signifikans.

Flere anerkendte måder at rapportere psykometriske resultater på anvendes i analyserne. Alle analyser anvender det traditionelle signifikansniveau ($p \leq .05$). Udtrykket standardforskel henviser til Glass Δ (Glass, McGaw, & Smith, 1981), som er sammenligneligt med Cohen's d (Cohen, 1988). De rapporterede værdier følger Cohen's anbefalinger og retningslinjer for tolkning af effektstørrelse (Cohen, 1988, 1992). Værdier for standardforskelle mellem .20 og .49 tolkes som små effektstørrelser, mens værdier mellem .50 og .79 tolkes som moderate og værdier fra .80 og op tolkes som store effektstørrelser. En anden måde at udtrykke effektstørrelse på er ved at præsentere dem i form af størrelsen af en gruppe som overlapper med en anden gruppe. En effektstørrelse på .20 svarer til et overlap på 92 %, .50 svarer til et overlap på 80%, mens .80 svarer til et overlap på 69 %. Jo større procentvise overlap jo mindre forskel er der på de to grupper. Cohen's d anvendes først og fremmest ved sammenligning af to grupperes observerede gennemsnit, mens Glass Δ anvendes, når en observeret gruppes gennemsnit sammenlignes med en teoretisk (forventet) værdi.

Den danske valideringsgruppe

Ved planlægningen af den danske dataindsamling blev det besluttet at tage udgangspunkt i de samme aldersintervaller, som anvendes i de originale amerikanske normtabeller. Af ressource-mæssige grunde blev det besluttet kun at planlægge en afprøvning af D-KEFS blandt 120 voksne personer i alderen 18–69 år.

Tabel 11.1 Rekrutteringsplan for D-KEFS valideringsundersøgelsen

Alder	N
18–19 år	30
20–29 år	15
30–39 år	15
40–49 år	15
50–59 år	15
60–69 år	30
N	120

Tabel 11.1 viser rekrutteringsplanen for dataindsamlingen. Der tilstræbes en ligelig fordeling af mænd og kvinder inden for hver aldersgruppe. Som det fremgår, er der dobbelt så mange personer i den yngste og den ældste aldersgruppe. Dette blev besluttet som en ekstra foranstaltning i tilfælde af, at det blev nødvendigt at anvende mere avancerede psykometriske metoder i analyserne, hvor flere aldersgrupper skulle slås sammen efter continuous norming metoden.

På baggrund af denne plan blev et tilfældig udtræk bestilt fra CPR-registret. Det bestilte udtræk omfattede 3 600 personer stratificeret efter alder og køn, således at hver celle i rekrutteringsplanen var overrepræsenteret med en faktor 30. Det bestilte udtræk bestod af personer bosiddende i og omkring København, uagtet statsborgerskab.

Valideringsundersøgelsen forløb i oktober og november 2011, og rekruttering af deltagerne forløb over 7 omgange i løbet af dataindsamlingsperioden. For hver omgang blev det vurderet, om der manglede testpersoner i hver aldersgruppe (hhv. mænd og kvinder), og et pseudotilfældigt udvalg af mænd og kvinder blev udtrukket i hver relevant aldersgruppe. Alder og køn har således fungeret som inklusionskriterier. Telefonnummer blev søgt på hver person og et introduktionsbrev blev sendt til vedkommendes adresse. Derefter blev personen ringet op og søgt rekrutteret til en D-KEFS-test. Alle personer med et fundet telefonnummer blev søgt kontaktet mindst 4 gange og på forskellige tidspunkter (arbejdstid, aften, hverdag og weekend).

Ved telefonrekrutteringen blev flg. eksklusionskriterier anvendt:

- Ikke-korrigeret høretab.
- Ikke-korrigeret synsnedsettelse.
- Motoriske handicap i overkroppen, som kunne hæmme testpersonens håndtering af testmaterialer.
- Utilstrækkelige dansk kundskaber. I praksis vurderedes disse som tilstrækkelige, hvis det var muligt at gennemføre telefonsamtalen med personen.
- Har eller har haft blodprop eller blødning i hjernen.
- Har eller har haft meningitis.
- Har eller har haft epilepsi.
- Har eller har haft skizofreni.
- Har eller har haft bipolar affektiv sindslidelse.

- Er eller har været misbruger af alkohol eller narkotika.
- Har været bevidstløs i mere end 5 minutter.
- Har været indlagt på sygehus på grund af en hovedskade.
- Har lidt af hukommelsesbesvær eller problemer med at koncentrere sig.

Ved telefonrekrutteringen blev testpersonernes uddannelse også registreret, men testpersonerne blev ikke udvalgt efter uddannelse. Bortfaldet blandt testpersonerne i rekrutteringsprocessen beskrives i tabel 11.2.

Tabel 11.2 Bortfald af testpersoner

Alder	Fra CPR	Søgt tlf.	Fundet tlf. (% ift. søgt)	Sendt brev	Deltagere (% ift. breve)
18–19 år	900	357	218 (61.1 %)	205	30 (14.6 %)
20–29 år	450	364	182 (50.0 %)	97	15 (15.5 %)
30–39 år	450	205	112 (54.6 %)	91	15 (16.5 %)
40–49 år	450	225	157 (69.8 %)	156	11 (7.1 %)
50–59 år	450	150	94 (62.7 %)	75	16 (21.3 %)
60–69 år	900	341	252 (73.9 %)	247	24 (9.7 %)

I løbet af dataindsamlingsperioden blev i alt 111 personer i alderen 18–69 år rekrutteret og testet med den danske version af D-KEFS. I alt deltog 57 mænd og 54 kvinder i dataindsamlingen (se tabel 11.3). Fordelingen af testpersoner efter køn afviger ikke signifikant fra det forventede ($\chi^2(5) = 5.714; ns$).

Tabel 11.3 Kønsfordeling i efter aldersgruppe

	Mænd	Kvinder
18–19 år	16 (53.3 %)	14 (46.7 %)
20–29 år	8 (53.3 %)	7 (46.7 %)
30–39 år	9 (60.0 %)	6 (40.0 %)
40–49 år	4 (35.4 %)	7 (63.6 %)
50–59 år	6 (37.5 %)	10 (62.5 %)
60–69 år	14 (58.3 %)	10 (41.7 %)
N	57 (51.4 %)	54 (48.6 %)

Fordelingen af personer efter uddannelsesniveau blev også sammenlignet med den forventede fordeling (jf. Danmarks Statistik, 2009; se tabel 12.4). Fordelingen af uddannelsesniveauer i den opnåede valideringsgruppe viser, at personer med de korteste uddannelser (U1) er underrepræsenterede i alle aldersgrupper sammenlignet med de forventede fordelinger i hver aldersgruppe i valideringsgruppen, hvis fordelingen havde svaret til normalbefolkningens. Og personer med uddannelser i gruppen U2 er underrepræsenterede i aldersgrupperne fra 30 til 69 år.

Tabel 11.4 Fordeling af uddannelsesniveauer i den opnåede valideringsgruppe (forventede fordeling jf. oprindelige plan i parentes)

Alder	U1*	U2*	U3*	U4*	N
18–19 år	13 (28)	16 (2)	0 (0)	0 (0)	29 (30)
20–29 år	1 (4)	10 (8)	3 (2)	1 (1)	15 (15)
30–39 år	1 (3)	3 (7)	7 (4)	4 (1)	15 (15)
40–49 år	0 (3)	3 (7)	7 (4)	1 (1)	10 (15)
50–59 år	3 (4)	4 (7)	8 (3)	1 (1)	16 (15)
60–69 år	3 (11)	7 (12)	10 (5)	4 (2)	25 (30)
N	21 (53)	43 (43)	35 (18)	11 (6)	110 (120)

* Længste afsluttede uddannelse: U1 = Folkeskole eller tilsvarende, U2 = Gymnasium/ Faglært eller tilsvarende, U3 = Kortere/mellemlang videregående (1–4 år) og U4 = Lang/forskeruddannelse (5+ år). 1 person er udeladt grundet manglende oplysning om uddannelse.

Valg af mål til analyser

D-KEFS omfatter ni test, og en komplet udredning med samtlige test giver mulighed for at beregne i alt 90 standardiserede mål med skalascorer ($M = 10$; $SD = 3$). Disse 90 mål er opdelt i hhv. 42 primære mål og 48 supplerende mål. Inden for de primære mål opdeles disse i 26 primære mål, 6 kombinationsmål og 10 kontrastmål. En oversigt over primære og supplerende mål findes i tabel 11.5.

Tabel 11.5 Kategorisering af standardiserede D-KEFS mål

Test	Primære	Kombination	Kontrast	Supplerende	Totalt
Trail Making	5	1	5	1	12
Ordmobilisering	4	2	–	9	15
Figurdesign	4	1	1	4	10
Farve-ord interferens	4	1	3	4	12
Sortering	3	1	1	20	25
Tyve spørgsmål	3	–	–	–	3
Begrebsforståelse	1	–	–	2	3
Tårnet	1	–	–	4	5
Talemåder	1	–	–	4	5
Total	26	6	10	48	90

Af de 42 primære mål defineres de 17 som kernemål ("key D-KEFS variables"), der har indgået i reliabilitetsberegningerne og beregning af *SEM* (Standard Error of Measurement) i det amerikanske standardiseringsarbejde (Delis et al., 2001b). Se tabel 11.6.

Tabel 11.6 D-KEFS kernemål

Test	Mål
Trail Making	Talrækkefølge og Bogstavorden kombinationsmål
Ordmobilisering	Fonemisk ordmobilisering, korrekte svar
Ordmobilisering	Semantisk ordmobilisering, korrekte svar
Ordmobilisering	Kategoriskift, korrekte svar
Ordmobilisering	Korrekte skift
Figurdesign	Sorte prikker korrekte figurer
Figurdesign	Kun hvide prikker korrekte figurer
Figurdesign	Skift korrekte figurer
Farve-ord interferens	Farvebenævnelse og Læsning kombinationsmål
Sortering	Bekræftede korrekte sorteringer totalt
Sortering	Beskrivelsesscore totalt Fri sortering
Sortering	Beskrivelsesscore totalt Forståelse af sortering
Tyve spørgsmål	Abstraktionsscore totalt
Tyve spørgsmål	Vægtede scorer totalt
Begrebsforståelse	Konsekutivt korrekte svar totalt
Tårnet	Nøjagtighed totalt
Talemåder	Total score Åbne spørgsmål

Gennemsnitspræstationer

Skalascorer ($M = 10$, $SD = 3$) anvendes i D-KEFS til at udtrykke en testpersons præstation. Den danske valideringsgruppes gennemsnitlige præstationer, udtrykt i skalascorer, blev sammenlignet med et forventet gennemsnit på 10 og en standardafvigelse på 3 svarende til det teoretiske gennemsnit i det amerikanske normgruppe. Alle følgende sammenligninger baseres på den samlede danske valideringsgruppe ($N = 111$).

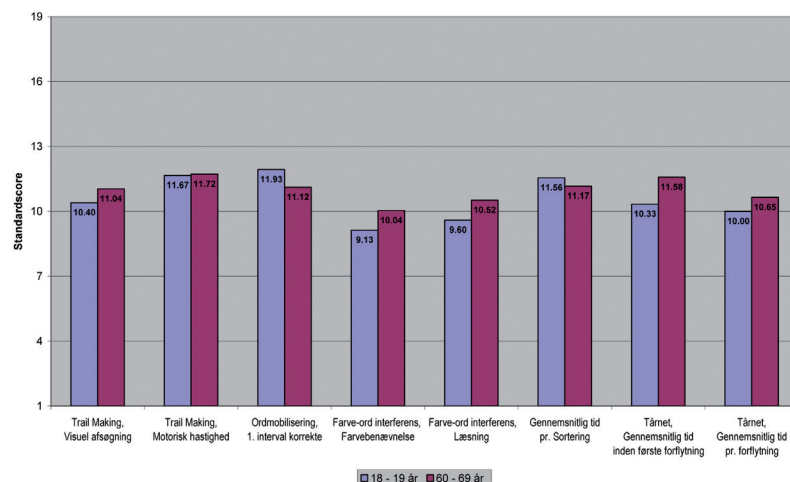
Generelt er scorerne i den danske valideringsgruppe højere end de forventede gennemsnit. Man kan antage, at udtræksplanens overvægt af 18–19 årige kan give forhøjede gennemsnitspræstationer i mål, hvor hastighed indgår. Dertil kan man antage, at underrepræsentationen af personer med de korteste uddannelser (og den tilsvarende overrepræsentation af personer med de længste uddannelser) generelt kan medføre relativt højere gennemsnitspræstationer i den danske valideringsgruppe.

For at afprøve dette blev følgende to hypoteser formuleret:

- Hypotese 1: Yngre præsterer bedre end ældre på mål, hvor hastighed er en vigtig del af præstationen.
- Hypotese 2: Personer med længere uddannelse præsterer generelt bedre end personer med kortere uddannelser.

Effekt af alder

Hypotese 1 blev afprøvet ved en en-vejs ANOVA (Bonferroni post hoc) mellem alle aldersgrupper og mål, hvori hastighed indgår. Denne analyse støttede ikke hypotesen om, at yngre personer præsterer bedre end ældre personer på mål hvor hastighed indgår som en vigtig del af præstationen. Figur 11.2 illustrerer dette ved en sammenligning af den yngste og den ældste aldersgruppe.



Figur 11.2 Sammenligning af gennemsnitsscorer ved mål, hvor hastighed indgår som en vigtig del af præstationen, for aldersgrupperne 18–19 år og 60–69 år.

Effekten af uddannelse

Hypotese 2 blev afprøvet ved først at klassificere den gennemsnitlige præstation for gruppen med uddannelsesniveau 4 (U4) for alle 90 standardiserede mål. Klassifikationen omfattede tre grupper; h = U4 har højeste gennemsnit (55); l = U4 har laveste gennemsnit (20) og * = U4 har hverken højeste eller laveste gennemsnit. Resultatet af klassifikationen blev anvendt som input

til en sign-test efter følgende formel (Siegler & Castellan Jr., 1988):

$$Z = \frac{|(h) - (l)| - 1}{\sqrt{(h) + (l)}}$$

I analysen fremkom et systematisk mønster, som var signifikant ($Z = 4.661$; $p < .00001$). Resultatet viser, at personer fra uddannelsesgruppe 4 har højere gennemsnitlige præstationer sammenlignet med personer fra de øvrige uddannelsesgrupper. En nærmere analyse blev foretaget med en en-vejs ANOVA (Bonferroni post hoc). Formålet var parvist at sammenligne hhv. U4 med U1 og U4 med U2 over samtlige (90) standardiserede mål. I alt 180 sammenligninger blev foretaget. Ud fra ren tilfældighed forventes ni af disse sammenligninger at falde ud som signifikante (5 %; $\alpha = .5$), men i analysen faldt ikke færre end 29 sammenligninger ud som signifikante (16 %). Forskellen mellem det observerede og det forventede antal signifikante forskelle er signifikant (test af to proportioner; $p < .00001$; Körner & Wahlgren, 1998). De to analyser understøtter hypotesen om, at personer med længere uddannelse generelt præsterer bedre end personer med kortere uddannelse. På denne baggrund og for at øge sammenligneligheden med den amerikanske normgruppe blev en tilfældig undergruppe udtrukket af den danske valideringsgruppe, hvor fordelingen af uddannelsesgrupper var repræsentativ for befolkningen i øvrigt.

Denne referencegruppe fandtes ved at bevare gruppen af testpersoner med de korteste uddannelser intakt ($n = 21$). Denne gruppe repræsenterede de 28 % af befolkningen (i de undersøgte aldersgrupper) med korteste uddannelser. Baseret på denne andel blev personer fra de øvrige uddannelsesgrupper trukket fra den opnåede valideringsgruppe i forhold til den oprindelige plan. Sammenlagt indgik 75 personer i referencegruppen fordelt over uddannelsesgrupperne 1 ($n = 21$), 2 ($n = 33$), 3 ($n = 15$) og 4 ($n = 6$).

Det første udgangspunkt for den grundlæggende vurdering af, om de amerikanske normer kan anses for egnede til brug ved danske testpersoner, har været, at gennemsnitspræstationen i den danske valideringsgruppe skal ligge inden for grænserne af et 95 % konfidensinterval omkring det forventede gennemsnit på 10 for hvert D-KEFS kernemål

Målingens standardfejl og konfidensinterval

Målingens standardfejl, eller målefejlen (*SEM*), giver indtryk af fejlstørrelsen i en observeret testscore. Eftersom *SEM* er omvendt relateret til reliabiliteten, så mindskes *SEM*, når reliabiliteten øges. Standardfejlen udtrykkes normalt i standard-score-enheder. *SEM* beregnes efter formlen:

$$SEM = SD\sqrt{1 - r_{xx}}$$

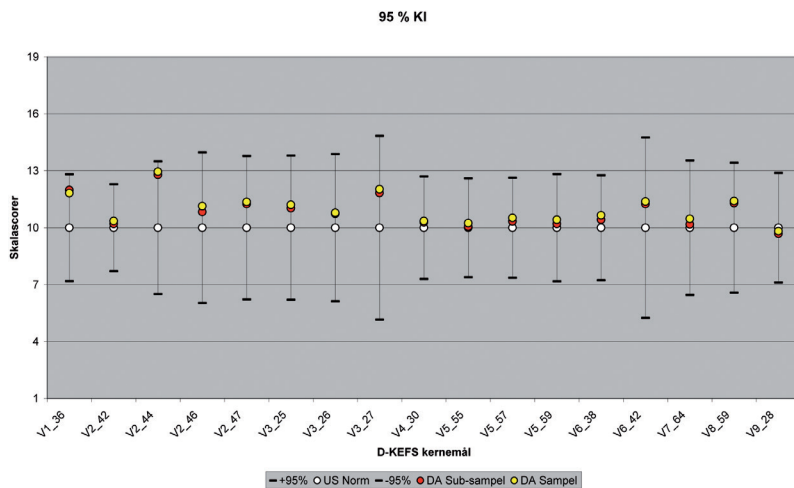
hvor *SEM* repræsenterer målingens standardfejl, *SD* er den observerede standardafvigelse og r_{xx} reliabiliteten. I målefejlen indgår tilfældige variationer, som kan bero på administreringsfejl eller gæt samt reliabilitetsproblemer (Magnusson, 2003).

SEM anvendes blandt andet til at beregne konfidensintervallet, det scoreinterval, som ligger omkring en observeret score og definerer området inden for hvilket den sande score forventes at findes. Konfidensintervaller udtrykker en scores præcision og tjener som påmindelse om, at målefejl indgår i alle testscorer, og at den observerede testscore kun er et estimat af den sande værdi.

Følgende formel anvendes for at konstruere konfidensintervallet for en given observeret skalascore:

$$p \text{ \% Konfidensinterval} = \text{Observeret skalascore} \pm zp^*(SEM)$$

hvor *p* er konfidensniveauet (oftest 90 % eller 95 %) og *zp* er *z*-værdien koblet til konfidensniveauet (1.645 for 90 % og 1.96 for 95 %). I figur 11.3 vises resultatet af sammenligninger mellem den gennemsnitlige præstation i hhv. den samlede danske valideringsgruppe og referencegruppen i forhold til den forventede gennemsnitlige præstation (skalascore 10). Konfidensintervallet på 95 % er markeret for hvert af de 17 D-KEFS kernemål.



V1_36 = Trail Making: Talrækkefølge og Bogstavorden kombinationsmål, V2_42 = Ordmobilisering: Fonemisk ordmobilisering korrekte svar, V2_44 = Ordmobilisering: Semantisk ordmobilisering korrekte svar, V2_46 = Ordmobilisering: Kategoriskift korrekte svar, V2_47 = Ordmobilisering: Korrekte skift, V3_25 = Figurdesign: Sorte prikker korrekte mønstre, V3_26 = Figurdesign: Kun hvide prikker korrekte, V3_27 = Figurdesign: Skift korrekte mønstre, V4_30 = Farve-ord interferens: Farvebenævnelse og Læsning kombinationsmål, V5_55 = Sortering: Bekræftede korrekte sorteringer totalt, V5_57 = Sortering: Beskrivelsesscore Fri sortering, V5_59 = Sortering: Beskrivelsesscore Forståelse af sortering, V6_38 = Tyve spørgsmål: Abstraktionsscore totalt, V6_42 = Tyve spørgsmål: Vægtede scorer totalt, V7_64 = Begrebsforståelse: Konsekutivt korrekte svar totalt, V8_59 = Tårnet: Nøjagtighed totalt, V9_28 = Talemåder: Total score.

Figur 11.3 Den samlede danske valideringsgruppe og referencegruppens gennemsnitlige præstationer ift. det teoretiske gennemsnit 10, omgivet af 95 % konfidensintervaller.

Det fremgår af figuren, at de gennemsnitlige præstationer i den samlede danske valideringsgruppe (markeret med gult) falder inden for grænserne af konfidensintervallerne for samtlige mål. Som forventet, pga. overrepræsentationen af højtuddannede personer, er der en generel tendens til at de gennemsnitlige præstationer ligger noget over det teoretiske gennemsnit på 10. Dette er tydeligst for Semantisk ordmobilisering (V2_44). Det fremgår yderligere, at referencegruppens gennemsnitlige præstationer (markeret med rødt) generelt ligger lidt lavere end de gennemsnitlige præstationer for den samlede valideringsgruppe (15 af 17: $p < .001$). Dette støtter, at uddannelsesniveaet har en vis betydning for præstationen i de 17 D-KEFS kernemål.

Med undtagelse af Semantisk ordmobilisering er effektstør-

relserne, dvs. forskellene mellem den samlede danske valideringsgruppes gennemsnitlige præstationer og det forventede gennemsnit på 10 små til moderate. For referencegruppen bliver forskellene endnu mindre, idet gruppens gennemsnitlige præstationer ligger nærmere det forventede gennemsnit på 10 end den samlede danske valideringsgruppe gør (men man bemærker, at til trods for den systematiske forskel i testpræstation efter uddannelsesniveau, så forekommer forskellene i reel præstation mellem referencegruppen og den samlede danske valideringsgruppe at være relativt små). Målet for Semantisk ordmobilisering udviser størst effekt i både den danske valideringsgruppe og i referencegruppen. Kun to sammenligninger (målene Trail Making: Talrækkefølge + Bogstavorden kombinationsmål og Semantisk ordmobilisering) går over grænsen for en moderat effektstørrelse (Glass $\Delta \geq .50$). I tabel 11.7 præsenteres den danske valideringsgruppes ($N = 111$) og referencegruppens ($N = 75$) gennemsnit og standardafvigelser, udtrykt i skalascorer, og effektstørrelser (Glass Δ) sammen med gennemsnit og standardafvigelser fra den amerikanske normgruppe ($N = 874$).

Tabel 11.7 Gennemsnit og standardafvigelse, som skal scores, samt effektstørrelser for den samlede danske valideringsgruppe og for referencegruppen, samt gennemsnit og standardafvigelse for den amerikanske normgruppe

Test	M (US)	SD (US)	M (Da)	SD (Da)	Glass Δ^*	M (Reference-gruppe)	SD (Reference-gruppe)	Glass Δ^*
Trail Making								
Talrækkefølge og Bogstavorden kombinationsmål	10.1	3.1	11.8	2.2	.61	12.0	2.4	.66
Ordmobilisering								
Fonemisk ordmobilisering korrekte svar	10.0	3.3	10.4	3.6	.12	10.2	3.6	.07
Semantisk ordmobilisering korrekte svar	10.0	3.1	13.0	3.5	.98	12.8	3.3	.93
Kategoriskift korrekte svar	10.0	3.4	11.1	3.2	.38	10.8	3.2	.28
Korrekte skift	10.4	3.2	11.4	3.1	.45	11.3	2.8	.42
Figurdesign								
Sorte prikker korrekte mønstre	10.2	2.9	11.2	2.7	.40	11.0	2.7	.34
Kun hvide prikker korrekte mønstre	10.1	2.9	10.8	2.4	.26	10.7	2.5	.24
Skift korrekte mønstre	10.1	3.0	12.0	2.8	.68	11.8	2.6	.61
Farve-ord interferens								
Farvebenævnelse og Læsning kombinationsmål	10.2	2.7	10.4	1.8	.12	10.3	1.8	.09
Sortering								
Bekræftede korrekte sorteringer totalt	10.5	2.8	10.2	2.7	.08	10.1	2.6	.02
Beskrivelsesscore Fri sortering	10.3	2.8	10.5	2.8	.17	10.3	2.7	.11
Beskrivelsesscore Forståelse af sortering	10.3	3.2	10.4	3.4	.14	10.2	3.3	.07
Tyve spørgsmål								
Abstraktionsscore totalt	10.0	3.0	10.6	2.6	.22	10.4	2.5	.14
Vægtede scorer totalt	10.1	2.9	11.4	2.7	.46	11.3	3.0	.42
Begrebsforståelse								
Konsekutivt korrekte svar totalt	10.0	3.0	10.5	2.5	.16	10.2	2.5	.06
Tårnet								
Nøjagtighed totalt	10.2	3.1	11.4	2.4	.47	11.3	2.5	.43
Tålemåder								
Total score, Åbne spørgsmål	10.0	2.9	9.8	2.7	.06	9.7	2.5	.11

* Anbefalede størrelser til tolkning af Glass $\Delta < .20$ ingen effekt; $.20-.49$ lille effekt; $.50-.79$ moderat effekt; $\geq .80$ stor effekt.

Reliabilitet

Reliabilitet er et udtryk for en målings præcision og dens stabilitet over tid og i forskellige situationer (Anastasi & Urbina, 1997; Sattler, 2008). Inden for klassisk testteori antages det at en observeret testscore er en approksimation af et individs hypotetiske sande score, det vil sige den score, som han eller hun ville have opnået, hvis testen var fuldstændig reliabel. Forskellen mellem den hypotetiske testscore og den opnåede testscore er målefejlen. En reliabel test har en relativt lille målefejl og giver et stabilt testresultat ved hver test. Ved tolkning af en testscore skal man altid tage testens reliabilitet i betragtning.

Reliabiliteten i en test kan beregnes på forskellige måder. Uanset metoden, så gælder det, at jo højere værdi (reliabiliteten kan variere mellem 0 og 1), desto højere grad af sikkerhed er der i et testresultat. Blandt forskellige mulige beregningsmetoder, som kunne anvendes, er test-retest reliabilitet fortrinsvis anvendt i D-KEFS. Eftersom den danske valideringsundersøgelse ikke omfattede re-testning, så kan test-retest reliabiliteten for de 17 kernemål ikke beregnes.

Validitet

Intern struktur

Korrelationerne mellem målene i en test giver et indtryk af testens interne struktur, det vil sige en indikation af, i hvilken grad skalaerne ”hænger sammen” og har relation til konstruktet, som tolkningen af testresultater hviler på. Som tidligere nævnt er målene i D-KEFS grupperet i fire forskellige kategorier inden for de ni test (se tabel 11.5). Mål inden for samme test kan have forskelligt sigte og behøver derfor ikke udvise stor sammenhæng med hinanden (fx i korrelationer, hvor den ene variabel udgøres af et kontrastmål). I sådanne tilfælde er det heller ikke usædvanligt at en korrelation er negativ.

Korrelationsmønstre: niveau og retning

Formålet med en sammenligning af korrelationsmønstrene blandt de forskellige D-KEFS mål i den danske valideringsgruppe og den amerikanske normgruppe, er at analysere, om samme interne struktur i testene findes i begge grupper. Resultaterne fra den danske valideringsgruppe bør være i overensstemmelse med den interne struktur, som definerer testens validitet. De 42 primære mål indgår i denne analyse

Det andet udgangspunkt for den grundlæggende vurdering af, om de amerikanske normer er egnede til brug ved danske testpersoner, har været, at korrelationer mellem målene i D-KEFS-testene skal være i paritet med de amerikanske, både hvad angår retning og niveau.

Afprøvningen af dette udgangspunkt blev gennemført som en signifikanstest af forskelle imellem korrelation efter følgende formel (Angsmark, 1968):

$$Z = \frac{\hat{r}_{US} - \hat{r}_{Da} + .5 \cdot \left(\bar{r}_{(US:Da)} \cdot \frac{n_{US} - n_{Da}}{n_{US} \cdot n_{Da}} \right)}{\sqrt{\frac{1}{n_{US} - 3} + \frac{1}{n_{Da} - 3}}}$$

hvor \hat{r}_{US} og \hat{r}_{Da} er Fisher's z-transformerede korrelationer og $\bar{r}_{(US:Da)}$ er Fisher's z-transformerede middelnkorrelation.

For at undersøge, om den interne struktur i den danske version af D-KEFS adskiller sig fra den amerikanske, blev der genereret korrelationsmatricer for alle primære mål (primære mål, kombinationsmål og kontrastmål) ved alle test bortset fra Begrebsforståelse, Tårnet og Talemåder, fordi disse test kun indeholdt ét primært mål hver. Interkorrelationerne er beregnet for den danske valideringsgruppe og vises over diagonalen i tabel 11.8 til 11.13. Korrelationernes retning og størrelse sammenlignes med de tilsvarende amerikanske værdier (under diagonalen i tabellerne).

Table 11.8 Interkorrelationer mellem mål i Trail Making. Dansk valideringsgruppe (N = 111) over diagonalen, amerikansk normgruppe (N = 874) under diagonalen

	VI_30	VI_31	VI_32	VI_33	VI_34	VI_36	VI_38	VI_40	VI_42	VI_44	VI_46
PrimærtVI_30	–	0.38	0.40	0.23	0.24	0.43	-0.53	-0.10	-0.18	-0.20	0.07
PrimærtVI_31	0.37	–	0.43	0.47	0.16	0.82	0.11	-0.44	0.02	-0.35	0.31*
PrimærtVI_32	0.35	0.57	–	0.53	0.16	0.86	0.17	0.15	-0.51	-0.33	0.38*
PrimærtVI_33	0.25	0.46	0.51	–	0.10	0.59	0.69	0.59	0.46	0.49	0.81
PrimærtVI_34	0.27	0.31	0.33	0.22	–	0.19	-0.07	-0.03	-0.05	-0.08	-0.49
KombinationsmålVI_36	0.41	0.89	0.88	0.54	0.36	–	0.18	-0.15	-0.31	-0.42	0.41*
KontrastVI_38	-0.57	0.08	0.14	0.61	-0.01	0.12	–	0.59	0.52	0.58	0.64
KontrastVI_40	-0.12	-0.50	-0.06	0.50	-0.06	-0.32	0.45	–	0.45	0.82	0.54
KontrastVI_42	-0.10	-0.08	-0.45	0.51	-0.08	-0.29	0.49	0.55	–	0.85	0.42
KontrastVI_44	-0.16	-0.42	-0.36	0.47	-0.11	-0.44	0.50	0.83	0.81	–	0.47
KontrastVI_46	-0.10	-0.08*	-0.45*	0.51	-0.08	-0.29*	0.49	0.55	1.00	0.81	–

*p < .05 (p < .001 Bonferronikorrektion for multiple sammenligninger)

VI_30 = Visuel afsøgning, VI_31 = Talrækkefølge, VI_32 = Bogstavorden, VI_33 = Skift, VI_34 = Motorisk hastighed, VI_36 = Talrækkefølge og Bogstavorden kombinationsmål, VI_38 = Skift kontra Visuel afsøgning, VI_40 = Skift kontra Talrækkefølge, VI_42 = Skift kontra Bogstavorden, VI_44 = Skift kontra Talrækkefølge + Bogstavorden kombinationsmål, VI_46 = Skift kontra Motorisk hastighed.

Tabel 11.9 Interkorrelationer mellem mål i Ordmobilisering. Dansk valideringsgruppe ($N = 111$) over diagonalen, amerikansk normgruppe ($N = 874$) under diagonalen

	V2_42	V2_44	V2_46	V2_47	V2_49	V2_51
PrimærtV2_42	–	0.56	0.32	0.27	0.49	-0.29
PrimærtV2_44	0.56	–	0.52	0.50	-0.45	-0.57
PrimærtV2_46	0.37	0.45	–	0.94	-0.18	0.41
PrimærtV2_47	0.28	0.35	0.75	–	-0.18	0.29
KontrastV2_49	0.52	-0.41	-0.06	-0.05	–	0.30
KontrastV2_51	-0.15	-0.47	0.58	0.42	0.31	–

V2_42 = Fonemisk ordmobilisering, V2_44 = Semantisk ordmobilisering, V2_46 = Kategoriskift korrekte svar, V2_47 = Korrekte skift, V2_49 = Fonemisk kontra semantisk ordmobilisering, V2_51 = Kategoriskift kontra Semantisk ordmobilisering.

Tabel 11.10 Interkorrelationer mellem mål i Figurdesign. Dansk valideringsgruppe ($N = 111$) over diagonalen, amerikansk normgruppe ($N = 874$) under diagonalen

	V3_25	V3_26	V3_27	V3_31	V3_29	V3_33
PrimærtV3_25	–	0.67	0.47	0.92	0.85	-0.34*
PrimærtV3_26	0.68	–	0.52	0.90	0.85	-0.27*
PrimærtV3_27	0.34	0.41	–	0.54	0.80	0.61*
Kombinationsmål V3_31	0.92	0.92	0.41	–	0.93	-0.33*
PrimærtV3_29	0.83	0.86	0.72	0.91	–	0.02
KontrastV3_33	0.54*	0.49*	-0.50*	0.57*	0.21	–

* $p < .05$ ($p < .003$ Bonferronikorrektion for multiple sammenligninger)

V3_25 = Sorte prikker, V3_26 = Kun hvide prikker, V3_27 = Skift, V3_31 = Sorte prikker og Kun hvide prikker kombinationsmål, V3_29 = Sorte prikker og Kun hvide prikker kombinationsmål, V3_33 = Skift kontra Sorte prikker og Kun hvide prikker kombinationsmål.

Tabel 11.11 Interkorrelationer mellem mål i Farve-ord interferens. Dansk valideringsgruppe ($N = 111$) over diagonalen, amerikansk normgruppe ($N = 874$) under diagonalen

	V4_25	V4_26	V4_27	V4_28	V4_30	V4_32	V4_36	V4_34
PrimærtV4_25	–	0.53	0.59	0.51	0.87	-0.31	-0.06	-0.12
PrimærtV4_26	0.59	–	0.43	0.44	0.86	-0.02	0.04	-0.19
PrimærtV4_27	0.50	0.46	–	0.55	0.60	0.59	-0.45	0.15
PrimærtV4_28	0.44	0.43	0.59	–	0.54	0.14	0.50	0.73
KombinationsmålV4_30	0.89	0.88	0.54	0.49	–	-0.16	-0.03	-0.19
KontrastV4_32	-0.46	-0.12	0.53	0.18	-0.32	–	-0.46	0.30
KontrastV4_36	-0.08	-0.02	-0.46	0.44	-0.06	-0.39	–	0.61
KontrastV4_34	-0.38	-0.38	0.10	0.57	-0.44	0.48	0.52	–

V4_25 = Farvebenævnelse, V4_26 = Læsning, V4_27 = Inhibering, V4_28 = Inhibering og skift, V4_30 = Farvebenævnelse og Læsning kombinationsmål, V4_32 = Inhibering kontra Farvebenævnelse, V4_36 = Inhibering og skift kontra Inhibering, V4_34 = Inhibering kontra Farvebenævnelse og Læsning kombinationsmål.

Tabel 11.12 Interkorrelationer mellem mål i D-KEFS Sortering. Dansk valideringsgruppe ($N = 111$) over diagonalen, amerikansk normgruppe ($N = 874$) under diagonalen

	V5_55	V5_57	V5_59	V5_63
PrimærtV5_55	–	0.97	0.73	-0.06
PrimærtV5_57	0.95	–	0.75	-0.10
PrimærtV5_59	0.61	0.62	–	0.58
KontrastV5_63	-0.26	-0.28	0.56	–

V5_55 = Bekræftede korrekte sorteringer, V5_57 = Beskrivelsesscore, Fri sortering, V5_59 = Beskrivelsesscore, Forståelse af sortering, V5_63 = Forståelse af sortering kontra Beskrivelsesscore, Fri sortering.

Tabel 11.13 Interkorrelationer mellem mål i Tyve Spørgsmål. Dansk valideringsgruppe ($N = 111$) over diagonalen, amerikansk normgruppe ($N = 874$) under diagonalen

	V6_40	V6_38	V6_42
PrimærtV6_40	–	0.22	0.75
PrimærtV6_38	0.28	–	0.30
PrimærtV6_42	0.77	0.23	–

V6_40 = Stillede spørgsmål totalt, V6_38 = Abstraktionsscore totalt, V6_42 = Vægtede scorer totalt.

Af de 122 interkorrelationspar, som indgik i analysen, opfyldte 115 (94.3 %) begge kriterier (samme fortegn og samme niveau); hvilket vi tolker som mangel på signifikant forskel mellem de sammenlignede korrelationer. De syv korrelationspar, som ikke opfyldte kriterierne inkluderer to specifikke kontrastmål – tre korrelationer med målet Skift kontra Motorisk hastighed (i Trail Making) og fire korrelationer med målet Skift kontra Sorte prikker + Kun hvide prikker kombinationsmål (i Figurdesign). Det er værd at bemærke, at de negative interkorrelationer kun findes i forbindelse med kontrastmål.

Sammenfatning

De psykometriske analyser af anvendeligheden af de amerikanske normer sammen med den danske version af D-KEFS har søgt at besvare følgende spørgsmål:

- Forekommer der forskelle, der kan tolkes som sande forskelle mellem de to populationer (USA og Danmark)?
- Har disse forskelle praktisk klinisk betydning?

Disse spørgsmål er søgt besvaret med flere analysemetoder og ud fra forskelligt perspektiv baseret dels på sammensætningen af referencegruppen og sammensætningen af den opnåede valideringsgruppe. Med nogle få undtagelser viser analyserne, at de danske gennemsnitlige præstationer ligger nær standardskalscoren 10 i de amerikanske normer, og at samtlige danske gennemsnitlige præstationer falder inden for et 95 % konfidens-

interval omkring skalascoren 10 for de 17 kernemål. For disse 17 mål nærmer de danske middelprestationer sig endnu mere den teoretiske middelværdi på 10, når der korrigeres for skævhed i uddannelsesniveau i det danske sample. De faktisk forekommende forskelle, baseret på den danske referencegruppe i forhold til de amerikanske normer, viser ingen eller lille – og i to tilfælde moderat – effekt (Glass Δ). At korrelationsanalysen, med få undtagelser (syv korrelationspar), har fundet samme mønster af korrelationer mellem primære mål som i den amerikanske normgruppe giver yderligere støtte til at anvende de amerikanske normer med den danske D-KEFS.

Sammenfattende understøtter resultaterne af analyserne af data fra den danske valideringsundersøgelse anvendeligheden af de amerikanske normer med den danske D-KEFS, idet de fundne forskelle ikke vurderes at have praktisk betydning. Det skal dog understreges, at den beskrevne undersøgelse er en valideringsundersøgelse som kun har omfattet voksne testpersoner på trods af, at D-KEFS-testene anvendes fra 8 til 89 år. Man bør derfor altid tolke resultater opnået med D-KEFS testene hos børn og under 18 år med en vis forsigtighed, og testlederen bør altid trække på sin egen kliniske erfaring ved tolkning og anvendelse af resultater fra testene.