NASJONALE UTVALGSPRØVER I SKRIVING 2015

TEKNISK RAPPORT

Datum: 10.03.2016

Författad av
G. B. Skar
J. M. Iversen

NTNU
Skrivesenteret
Nasjonalt senter for skriveopplæring og skriveforsking
Skriveprøven

Förord

Denna tekniska rapport är skriven av Gustaf B. Skar och Jon Marius Iversen vid Nasjonalt senteret for skriveopplæring og skriveforskning, NTNU. Skar är huvudförfattare och Iversen författare av kapitel om faktoranalyser.

Rapporten har kvalitetsgranskats av Utdanningsdirektoratet.

Innehåll

Förord 2
Innehåll 3
1 Inledning 4
1.1 Utvalgsprøven 4
1.1.1 Syfte och uppdragsgivare 4
1.1.2 Konstrukt och uppgifter 4
1.1.3 Bedömning 5
2 Material och materialinsamling 7
2.1 Materialinsamling: Fas 1 7
2.2 Materialinsamling: Fas 2 8
3 Psykometrisk analys 10
3.1 Antaganden 11
3.2 Mått på psykometrisk kvalitet 12
3.3 MFRM-resultaten och nationell nivå 13
4 Psykometrisk kvalitet 15
4.1 Faktoranalys 15
4.2 Fit, reliabilitet, och skalstruktur 18
4.2.1 Elever, uppgifter, bedömningsområden och provtillfälle 19
4.2.2 Bedömare 20
4.2.3 Skalstruktur 21
4.3 Sammanfattning – från observerad till rättvis poäng 22
5 Nationell nivå 24
Avslutning 33
Litteratur 35
Appendix 37

1 Inledning

I denna tekniska rapport beskrivs arbetet med att mäta elevers skrivförmåga inom ramen för *de nasjonale utvalgsprøvene i skriving* (hädanefter: utvalgsprøven). I rapporten redovisas också den ”nationella nivån”, dvs. resultat på olika nivåer av aggregering för respektive årstrinn. Rapporten är uppbyggd på följande sätt: inledningsvis avges en kortfattad presentation av utvalgsprøven, dess syfte och omfattning. Därefter presenteras materialet och metoder för datainsamling och dataanalys. Själva resultatdelen är uppdelad i två sektioner. I den första beskrivs resultaten av de psykometriska undersökningarna och i den andra den nationella nivån.

1.1 Utvalgsprøven

1.1.1 Syfte och uppdragsgivare

Syftet med utvalgsprøven är tvåfaldigt. För det första syftar provet till att mäta elevers skrivförmåga efter fyra respektive sju års skrivundervisning. Provet administreras därför varje höst till ett nationellt representativt urval elever i femte och åttonde årskurs. För det andra, och viktigast, syftar provet till att skapa underlag för den *læringsstøttende prøven i skriving* (hädanefter: Skriveprøven), genom att på en bred målgrupp pröva uppgifter och inhämta information om nationell nivå. Resultaten från utvalgsprøven används för att skapa jämförelseunderlag till skolor som senare genomför Skriveprøven.

Utvalgsprøven utarbetas av Skrivesenteret på uppdrag av Utdanningsdirektoratet. Direktoratet har i sin tur uppdragits av Kunnskapsdepartementet att i enlighet med Stortingsmedling 22, «Motivation – mestring – muligheter – ungdomstrinnet» (Kunnskapsdepartementet, 2011), utveckla skrivprov för 5. och 8. trinn. Direktoratets uppdrag regleras i *Oppdragsbrev 29-10*.

1.1.2 Konstrukt och uppgifter

Konstruktionen av utvalgsprøven bygger på den teoretiska förståelse av skrivande som uttrycks i det så kallade Skrivhjulet (Berge, Evensen, & Thygesen, n.d.). Enligt denna förståelse innebär det *att skriva* att med semiotiska redskap utföra en eller flera skrivhandlingar för att

uppnå ett eller annat kommunikativt syfte. Operationaliserat i utvalgsprøven innebär detta att uppgifterna konstrueras så att de inkluderar ett syfte, en mottagare, ett publikationssammanhang och en skrivhandling (se exempel i Skar, Evensen, & Iversen, 2015).

I regel är utveckling av provuppgifter en treårig process. Alla uppgiftsförslag screenas i en pre-pilotering innan de under innevarande eller nästkommande år prövas i en pilot på ett nationellt representativt urval elever. Året därpå ingår uppgifter som klarat sig så långt i utvalgsprøven. (Provutvecklingen finns närmare beskriven i Holten-Kvistad och Skar (manus)). Av dessa skäl förekommer få justeringar av uppgifter efter att de har givits som utvalsgprøve. Uppgifterna i det nedan rapporterade utvalgsprøven har analyserats i en tidigare rapport (Skar & Iversen, 2015).

Från och med 2015 skriver varje elev två uppgifter. Detta har att göra med en omfattande kunskap på skrivbedömningsfältet om att elevers prestationer varierar starkt mellan uppgifter (jfr Bouwer, Béguin, Sanders, & van den Bergh, 2015; Breland, Bridgeman, & Fowles, 1999), vilket innebär att lösning av två uppgifter leder till en sannare nationell nivå.

1.1.3 Bedömning

I utvalgsprøven sköts bedömningen av tränade bedömare från den nationella bedömarpanelen. I denna ingår lärare som tränats av Skrivesenteret under flera år. Panelen representerar hela landet och båda de aktuella trinnen.

För att bedöma elevtexter har panelen tillgång till exempeltexter och bedömningskriterier. I Skriveprøven kallas de senare för mestringsnivåbeskrivelser (hädanefter: MNB). För att efterlikna det övriga nationella proven i Norge finns två uppsättningar MNB, en för 5. trinn och en för 8. trinn. Bedömningen i utvalgsprøven är strikt analytisk och för varje trinn finns sex skalor. De sex bedömningsskalorna är: *kommunikation*, *innehåll*, *textstruktur*, *språkbruk*, *stavning* och *interpunktion*. De fyra första mäter funktionskompetenser som är specifika för uppgiften. De två sista mäter kodningskompetenser, som i mindre utsträckning är avhängiga den specifika uppgiftsbeställningen. Skalorna förkortas ofta V1–6, där *V* står för *vurderingsområde*. Skalorna på 5. trinn består av tre steg och skalorna på 8. trinn av fem steg.

På både 5. trinn och 8. trinn finns dessutom skalsteget 0, som betyder att texten av någon grund inte är bedömningsbar. Typiskt för att den är för kort.

De ursprungliga MNB har sin utgångspunkt i ett forskningsprojekt där man inhämtade kunskap om vilken skrivförmåga lärare förväntade sig att elever hade efter fyra respektive sju års adekvat skrivundervisning (Matre et al., 2011). Dessa ”förväntningsnormer” motsvarar det mittersta steget på respektive skala. Till skillnad från vad som är vanligt i andra sammanhang finns det inte specialskrivna MNB för olika målgrupper (Alderson, 1991), utan panelen arbetar med samma kriterier som lärare sedan använder sig av i Skriveprøven.

2 Material och materialinsamling

I enlighet med uppdraget genomförs utvalgsprøven på ett nationellt representativt urval elever. För utvalgsprøven 2015 (hädanefter: U2015) består materialet av bedömningar av 1.342 texter skrivna av 671 elever, som representerar 41 skolor och två trinn. Materialet finns närmare beskrivet i tabell 2.1

Tabell 2.1 Antal elever, texter, bedömningar och skolor.

		5. trinn		8. trinn	
		n	%	n	%
Elever	*N*	320	100	351	100
	Flickor	166	51,9	196	55,8
	Pojkar	154	48,1	155	44,2
	Bokmål	264	82,5	247	70,4
	Nynorsk	56	17,5	104	29,6
	L1	295	92,2	317	90,3
	L2	25	7,8	34	9,7
Texter	*N*	640	100	702	100
Bedömningar	*N*	8.550	100	11.040	100
Skolor	*N*	20	100	21	100

Notera. All demografisk data bygger på självrapportering. L1/L2 = norska som första eller andraspråk. Deltagarinformationen disaggregeras inte på skolnivå (de enheter som anges i tabellen representerar således lägsta graden av aggregering). Bedömning avser antalet bedömningspunkter (elever * texter * bedömningsskalor).

Som framgår av tabellen finns är flickor och pojkar någorlunda jämt fördelade på båda trinnen. Andelen elever som skriver på nynorsk är skiljer sig åt; på 5. trinn uppgår de till en knapp femtedel, medan de på 8. trinn uppgår till en knapp tredjedel. Andelen andraspråkstalare är jämt fördelade.

2.1 Materialinsamling: Fas 1

Materialet samlas in i två steg. I det första använder Skrivesenteret listor som Utdanningsdirektoratet har utarbetet över slumpvis valda skolor för att kontakta eventuellt deltagande skolor. Eftersom provgenomförande är frivilligt lämnar Skrivesenteret en förfrågan. Förfrågan lämnas på våren, eftersom provet skall genomföras i starten på hösten. För U2015 gällde det speciella att deltagande elever ombads att skriva två uppgifter, vilket

innebar att några skolor tackade nej till medverkan. När Skrivesenteret har fått positivt besked från minst 20 skolor anses antalet deltagare vara tillräckligt.

Under våren färdigställs också provmaterialet. För U2015 bestod detta av två uppgifter för respektive trinn. För femte trinn rörde det sig om en uppgift relaterad till skrivhandlingen *att föreställa sig* och en till *att argumentera*. För åttonde trinn rörde det sig om skrivhandlingarna *att föreställa sig* och *att utforska*. (Uppgifterna finns beskrivna i Skar & Iversen, 2015). Varje uppgift får en intern kod, som byggs upp på följande sätt: provtyp, år, trinn och skrivhandling. Utvalgsprøven '15, 5. trinn, att föreställa sig blir således *U1556*. Se vidare tabell 2.2.

Tabell 2.2 Uppgifter och uppgiftskoder.

	5. trinn		**8. trinn**	
Skrivhandling	att föreställa sig	att överbevisa	att utforska	att föreställa sig
Kod	U1555	U1556	U1584	U1585

Notera. Skrivhandlingarna (och nummer) finns beskrivna i tidigare tekniska rapporter.

Skolor som deltar får genom ett tryckeri tillsänt provmaterialet från Skrivesenteret. Materialet innehåller information om uppgifter och hur provet skall administreras. Det är skolans lärare som sköter administrationen.[1] För att inte introducera onödig variation i resultaten ordnades materialet för U2015 så att hälften av de deltagande skolorna mottog paket där U1555 respektive U1585 skulle lösas först och hälften av skolorna paket där U1556 respektive U1584 skulle lösas först (jfr eng. *counterbalancing*). I materialet var detta ordnat så att uppgifterna namngavs som *Oppgave A* respektive *Oppgave B*.

När eleverna har genomfört provet skickas elevlösningar till Skrivesenteret. Där anonymiseras, kodas och skannas dem för digital lagring och distribution till bedömare. I samband med kodningen upprättas anonymiserade datamatriser där elevkoden endast kan användas till att särskilja elever, men inte för att spåra dem.

2.2 Materialinsamling: Fas 2

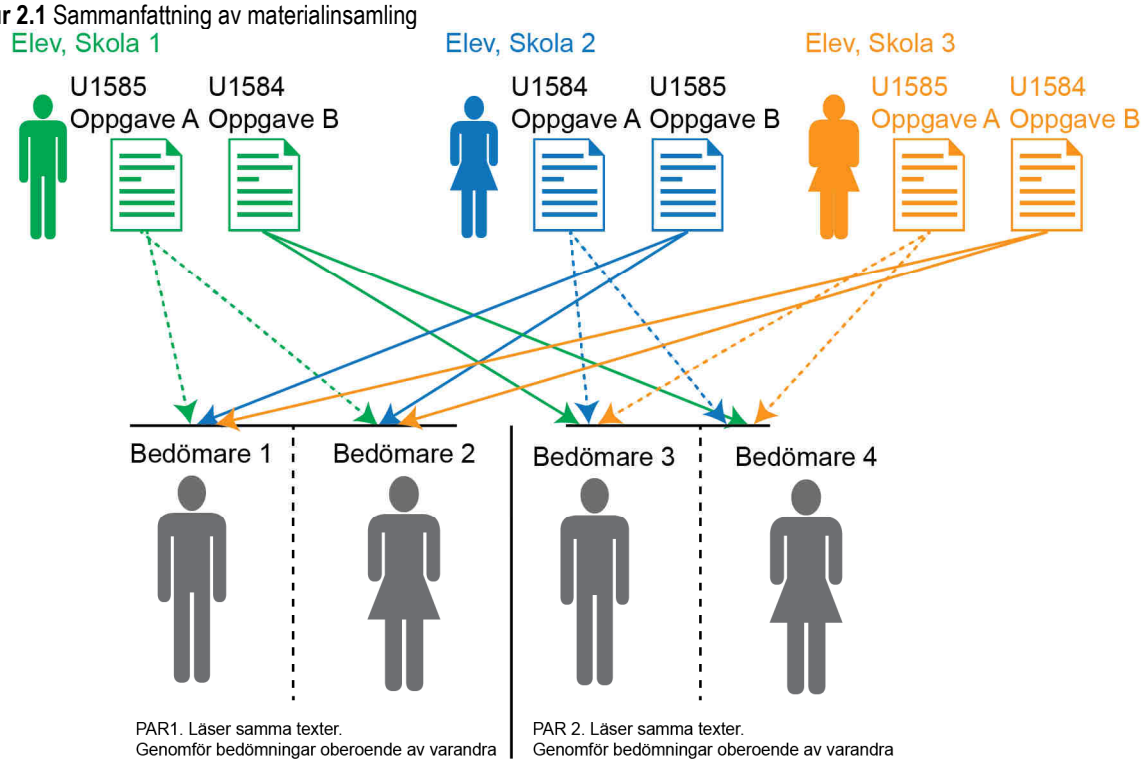
I materialinsamlingens andra steg distribueras elevtexter till den nationella panelen av expertbedömare. Distributionen sker genom ett krypterat filöverföringssystem tillhandahållet

[1] Skrivesenteret genomför varje år besök vid skolor där man administrerar pilotprov. Dessa besök har inte givit indikationer på validitetshotande variationer i sättet att administrera provet.

av NTNU och förser bedömare med anonymiserade elevtexter. Detta innebär att bedömaren varken har tillgång till uppgifter om elevens identitet eller demografisk bakgrundsinformation.

Varje text distribueras till minst två panelmedlemmar, som bedömer texterna enligt MNB och exempeltexter. (Av psykometriska skäl bedöms fyra texter av samtliga panelmedlemmar). Varje text bedöms på sex skalor och får totalt minst 12 omdömen. För U2015 gällde dessutom att ingen bedömare läste mer än en text från samma elev. Detta innebär att en elev får minst 24 oberoende bedömningar.[2] Bedömningarna registreras på en lösenordskyddad e-registreringstjänst utvecklad på NTNU (avd. HiST). I figur 2.1 sammanfattas denna och föregående materialinsamlingsfas.

Figur 2.1 Sammanfattning av materialinsamling



[2] Detta gäller för 669 av de totalt 671 eleverna. För de två övriga eleverna, som båda är elever på 5. trinn, saknas i det ena fallet bedömning på grund av blankt svar (av det skälet har U1555 färre observationer) och i det andra fallet på grund av en bedömare som inte sänt in fullständig data.

3 Psykometrisk analys

Bedömningen på utvalgsprøven är analytisk, men som nationell nivå redovisas sammanslagna poäng på en skala med 50,0 i genomsnitt. Det betyder att de 24 resultat som en elev erhåller reduceras till *ett* tal.

För att ta fram en psykometriskt hållbar nationell nivå analyseras resultaten på U2015 i huvudsak med hjälp av Rasch-analyser, som genomförts i datorprogrammet *Facets* (Linacre, 2014).[3] Dessa bygger i sin tur på att data passas till Rasch-modellen (Rasch, 1980), som i sin enklaste form gör gällande att ”sannolikheten för ett korrekt svar är en funktion av skillnaden mellan en testtagares förmåga och ett items svårighet” (Barkaoui, 2014, s. 1301). Rasch-modellen anger alltså att en testtagares resultat beror på elevens förmåga och uppgiftens svårighet och inget annat. En viktig förutsättning, vilken vi återkommer till, är därför att data passar modellen (och inte motsatt, som vanligen är fallet).

Den grundläggande Rasch-modellen är avpassad dikotoma items. Linacre (1989, citerad i Linacre, 2013) har utvidgat Rasch-modellen till den så kallade *many-facet reasch measurement*-modellen (MFRM).[4] I denna kan fler aspekter förklara en testtagares resultat, t.ex. bedömares stränghet.

Följande MFRM-modell (Barkaoui, 2014; Eckes, 2015; Linacre, 2013), som bygger på *Rating Scale-modellen*, är den som använts i U2015[5]:

$$ln\left[\frac{p_{niljk}}{p_{niljk-1}}\right] = \theta_n - \beta_i - \delta_l - \alpha_j - \tau_k, \qquad (1.1)$$

[3] En utförligare presentation av Rasch- och MFRM-analyser lämnas i Skar och Iversen (2015).

[4] Andra har också utvidgat modellen, t.ex. Andrich (1978) och Masters (1982), båda citerade i Eckes (2015).

[5] För U2015 har modellen byggts ut för att inkludera även andra aspekter, nämligen kön, målform, L1/L2 och tidpunkt för genomförande av uppgift. Dessa har inkluderats som s.k. dummy-facets för att möjliggöra interaktionsanalyser. Huvudförfattaren tillhandahåller fullständiga specifikationer vid förfrågan.

där

P_{niljk} är sannolikheten att testtagare *n*, på bedömningsområde *i*, på uppgift *l*, av bedömare *j* observeras i kategori *k*

$P_{niljk-1}$ är sannolikheten att testtagare *n*, på bedömningsområde *i*, på uppgift *l*, av bedömare *j* observeras i kategori *k-1*

θ_n är testtagarens skrivförmåga

β_i är kriteriets/bedömningsområdets svårighet

δ_l är uppgiftens svårighet

α_j är bedömarens stränghet och

τ_k är svårigheten att få ”k” (t.ex. ”3”) relativt till att få ”k-1” (t.ex. ”2”)

I Rasch-litteraturen kallas testtagare (eller personer), kriterier, uppgifter, bedömare för facets. En enskild elev, uppgift, bedömare och så vidare är ett *element* inom ett facets.

Om data passar MFRM-modellen kan analysen användas för att både undersöka den psykometriska kvaliteten vid mätningen av skrivförmåga och för att producera resultat. I de kommande tre avsnitten presenteras kortfattat grundläggande antaganden för användning av MFRM-modellen, mått på psykometrisk kvalitet och hur MFRM-resultaten använts för att producera den nationella nivån.

3.1 Antaganden

Tolkningen av resultaten från en MFRM-analys bygger på antaganden om s.k. lokalt oberoende (*local independence*) och endimensionalitet (Woods & Baker, 1985). Lokalt oberoende betyder att items inte är beroende av varandra. Ett konkret exempel är att möjligheten att svara rätt på en fråga inte är beroende av att ha svarat rätt på en annan fråga. Endimensionalitet avser ett krav om att items (eller bedömningsområden) har förmågan att forma ett ”entydigt” mönster av empiriska observationer (Eckes, 2015). Ett exempel är prov med många dikotoma items. Låt säga att 20 av 30 items bidrar till att mäta grammatisk kunskap, medan 10 items fokuserar på matematisk förmåga. Då omöjliggörs valid tolkning av det sammanfattande provresultatet och vi kommer inte att kunna göra meningsfulla analyser av vare sig enskilda items, deras relation eller provresultaten som helhet. Skulle vi däremot ta

bort de 10 items som relaterar till matematik återstår 20 item som på olika sätt bidrar till att mäta grammatikkunskap (se fler exempel i Bond & Fox, 2015). Psykometrisk endimensionalitet skall dock inte förstås som krav på att konstruktet som mäts behöver vara ”psykologiskt endimensionellt” (McNamara, 1996). Tvärtom vet vi exempelvis skrivförmåga kan förstås som flera, relaterade aspekter eller dimensioner.

Endimensionalitet kan skattas med exempelvis principal komponent-analys (PCA), eller annan faktoranalys, eller med så kallad fit-statistik. Skattningar av det senare rapporteras i *Facets* för element, facets och för data-setet som helhet. Det senare skattas med vad som kallas *global fit*, något som med ett tekniskt begrepp uttrycks i ”log-likelihood chi-square”. Som estimat på gobal fit är detta mått emellertid problematiskt, eftersom ett givet data-set nästan aldrig fullständigt passar modellen (se mer i Eckes, 2015). Ett bättre mått är därför det som kallas absoluta standardiserade residualer (ibid.).

Tumregeln säger att data passar modellen bra om bara ett visst procenttal av observationerna som används för estimering har en standardiserad residual som överstiger 2,0 respektive 3,0. Enligt Linacre (2013) passar data om sammantaget maximalt cirka 5 % av observationerna har en standardiserad residual över 2,0 och 1 % över 3,0.

För att undersöka om data passar MFRM-modellen har båda analyser genomförts. Dessa presenteras inledningsvis i kapitlet om psykometrisk kvalitet.

3.2 Mått på psykometrisk kvalitet

I Rasch-analyser skattas elementvärden i logaritmiska odds (s.k. *logits*). Transformeringen från rådata till logits innebär dels transformering från ordinal- till intervallskala, dels att alla facets placeras på samma skala (Bond & Fox, 2015; McNamara, 1996). För att kunna fästa tilltro till resultaten måste flera mått indikera kvalitet. I tabell 3.1 anges namnen på de viktigaste av dessa mått och vad de indikerar.

Tabell 3.1 Kvalitetsmått i MFRM-analyser

	Index	Typ av index/estimat	Förklaring
Alla facets	Q	Homogenitetsindex	Chi-square-statistik som testar antagandet om att det inte finns någon signifikant skillnad mellan element. Vid signifikans: åtminstone två element skiljer sig åt
	H	Separationsindex	Antalet statistiskt distinkta klasser inom ett facet
	R	Reliabilitetsindex	Separationsindexets reliabilitet (analogt med Cronbach alpha); kan anta värde upp till 1,0. Höga värden innebär trovärdighet i diskriminering mellan element.
	Infit	Passar data till modell	Skillnader mellan observerade och av modellen förväntade värden. Signifikant infit över 1,30 indikerar att data inte passar modellen. Signifikant infit under 0,75 indikerar att data passar modellen ”för väl”, t.ex. på grund av redudanta items, eller att bedömare använder få skalsteg.
	Outfit	Passar data till modell	Skillnader mellan observerade och av modellen förväntade värden. Känslig för uteliggare (outlier). Signifikant outfit över 1,30 indikerar att data inte passar modellen. Signifikant outfit under 0,75 indikerar att data passar modellen ”för väl”, t.ex. på grund av redudanta items, eller att bedömare använder få skalsteg.
Bedömare	Fullständig samstämmighet	Procent fullständig samstämmighet	-
	SR-ROR	Korrelation. (“Single Rater–Rest Of Raters”)	Ju högre korrelationen är, desto mer går bedömarna i samma riktning. Värden inom området ,30–,70 anses acceptabla. För hög korrelation indikerar att bedömarna inte beter sig enligt Rasch-modellen.
Skalor	Rasch-Andrich Thresholds		Värdet anger var på logitskalan som två kategorier är lika sannolika; skall öka med kategorier

Notera. För teknisk beskrivning av index, se exempelvis Eckes (2015), Knoch (2007) Linacre (2013), Myford och Wolfe (2003, 2004, 2009), Schumacker & Smith (2007) och Skar och Iversen (2015).

Ytterligare två kvalitetsindikatorer ingår i undersökningen. För det första tillåter *Facets* användaren att göra en grafisk inspektion av relationen data–modell och vi har tagit fram dessa grafer. För det andra har vi mätt bedömarsamstämmigheten också med *Intraclass Correlation Coefficient* (ICC).

3.3 MFRM-resultaten och nationell nivå

MFRM-analysen resulterar i estimat kopplade till varje element uttryckt i logits. Det betyder att elevers skrivförmåga skattas i logits, liksom görs bedömares grad av stränghet (vilket är den psykometriska termen), uppgifters grad av svårighet och bedömningsområdenas svårighet.

I enlighet med direktiv från Utdanningsdirektoratet transformeras elevers logits-värden till ett nytt skalpoäng, som representerar den sammanlagda förmågan med hänsyn tagen till båda uppgifter och alla sex bedömningsområden. Medelvärdet på den nya skalan är 50,0. (I kapitlet

om nationell nivå återges formeln för transformering.) För alla andra Facets behålls logit-värdena.

Facets transformerar också per automatik logit-värden tillbaka till den ursprungliga skalan. Dessa konverterade poäng kallas i *Facets*-terminologi, liksom de gör här, för *fair average*.

4 Psykometrisk kvalitet

Detta kapitel avhandlar den psykometriska kvaliteten relaterad till U2015. Inledningsvis kommenteras mått som rör grundantaganden för Rasch-modellen. Därefter presenteras resultaten av analyserna av kvalitetsindikatorerna, det vill säga värden för fit, reliabilitet, signifikans och skalstruktur. Vi kommenterar också skillnaderna i resultat mellan provtillfälle 1 och provtillfälle 2.

Som tidigare nämnt kan andelen standardiserade residualer användas för att skalla global fit. I tabell 4.1 redovisas antal och andel residualer som överstiger 3,0 respektive 2,0. Andelen som överstiger 3,0 skall helst inte vara mer än 1 % och andelen som överstriger 2,0 helst inte vara mer än 5 %.

Tabell 4.1 Standardiserade residualer

	Obs.	**Residual > 3,0**		**Residual > 2,0**	
		n	%	n	%
5. trinn	8.550	9	0,1	432	5,1
8. trinn	11.040	42	0,4	472	4,2

Notera. Obs = antal observationer.

Som vi kan se i tabellen indikerar resultaten att data på det stora hela passar modellen. I faktoranalysen, som redovisas i nästa avsnitt, får vi dock en mer nyanserad bild gällande antagandet om endimensionalitet.

4.1 Faktoranalys

Som ved tidligere gjennomganger skal vi analysere forholdet mellom vurderingsområdene. Vi analyserer faktorstrukturen for å se om de ulike vurderingsområdene lader opp mot en og samme faktor eller om det er indikasjoner på en tofaktorstruktur. Slik som ved gjennomgangen våren 2015 inkluderer vi først en enkel korrelasjonsmatrise (tabell 4.2). Vi ser at de fire første vurderingsområdene kommunikasjon, innhold, tekstoppbygging og språkbruk er relativt sterkt innbyrdes korrelerte, mens korrelasjonen er lavere mot rettskriving og tegnsetting. På den andre siden er tegnsetting og rettskriving innbyrdes relativt sterkt korrelerte. Dette er som tidligere og er en indikasjon på at fire av vurderingsområdene henger nærmere sammen enn de to andre, som på sin side har mye felles variasjon.

Tabell 4.2 Korrelasjonsmatrise av vurderingsområdene på 5.trinn

	Kommunikasjon	Innhold	Tekstoppbygging	Språkbruk	Rettskriving	Tegnsetting
Kommunikasjon	1					
Innhold	0,588	1				
Tekstoppbygging	0,609	0,606	1			
Språkbruk	0,595	0,595	0,615	1		
Rettskriving	0,360	0,348	0,365	0,442	1	
Tegnsetting	0,395	0,377	0,412	0,498	0,533	1

Tabell 4.3 angir en prinsipal komponentanalyse for det samme trinnet. Tidligere har vi sett at vi stort sett har hatt en en-faktormodell, hvor en andre faktor har vært inkludert, men med en eigenvalue like under 1. Det har blitt benyttet en verdi på 1 som vanlig cutoff, men faktoren har blitt inkludert for å illustrere de ulike vurderingsområdenes lading opp mot faktorene.

I analysene er det gjennomført *oblimin rotasjon*. I PCA-analyser gjennomføres rotasjon av faktordimensjonene funnet i det initielle uttaket av faktorer, for å kunne oppnå tolkbare resultater av analysen. Vi skiller da ofte mellom *orthogonal* og *oblique rotasjon*. Hovedforskjellen mellom disse er at orthogonal rotasjon antar at faktorene i analysen er ukorrelert, mens oblique rotasjon antar at faktorene er korrelert. Siden faktorene i vårt tilfelle har en relativt høy korrelasjon, velger vi å gjennomføre en oblique rotasjon. Som rotasjon i denne gjennomgangen brukes alternativet av oblique rotasjon som kalles *direct oblimin*.

I utgangspunktet resulterer prinsipal komponentanalysene i en faktor. En rotasjon av den ene faktoren gir ingen mening. Siden vi likevel har en annen faktor som er i nærheten av verdien en, velger vi å gjennomføre en rotasjon med to faktorer.

Tabell 4.3 Prinsipal komponentanalyse av vurderingsområdene på 5.trinn

	Faktor 1	Faktor 2
Eigenvalue	3,467(58 %)	0,894 (15 %)
Kommunikasjon	0,850	-0,023
Innhold	0,874	-0,063
Tekstoppbygging	0,854	-0,009
Språkbruk	0,705	0,217
Rettskriving	-0,042	0,904
Tegnsetting	0,067	0,827

Resultatet av analysen er som nevnt at den indikerer en enfaktormodell. Som ved flere av de tidligere gjennomgangene er det fire av vurderingsområdene, nemlig kommunikasjon, innhold, tekstoppbygging og språkbruk som lader opp mot denne faktoren. De to andre vurderingsområdene, rettskriving og tegnsetting, lader svært lite mot den ene faktoren. Den

andre faktoren påvirkes imidlertid sterkt av tegnsetting og rettskriving, men ikke av de fire andre faktorene. Faktorstrukturen ser derfor ut til å være veldig lik fra forrige gjennomgang. Faktor 1 forklarer 58 % av variansen, mens faktor 2 forklarer 15 %.

Også på 8. trinn har vi gjennomført samme analyser. Først ser vi fra korrelasjonsmatrisen i tabell 4.4 at samme mønster som på 5.trinn gjentar seg. De fire vurderingsområdene kommunikasjon, innhold, tekstoppbygging og språkbruk korrelerer sterkt innbyrdes. Korrelasjonen er også sterkere enn på 5. trinn. Disse fire vurderingsområdene korrelerer også her mindre med rettskriving og tegnsetting, mens disse igjen er relativt sterkt innbyrdes korrelerte.

Tabell 4.4 Korrelasjonsmatrise for vurderingsområdene på 8.trinn

	Kommunikasjon	**Innhold**	**Tekstoppbygging**	**Språkbruk**	**Rettskriving**	**Tegnsetting**
Kommunikasjn	1					
Innhold	0,77	1				
Tekstoppbygging	0,74	0,72	1			
Språkbruk	0,71	0,68	0,70	1		
Rettskriving	0,57	0,50	0,53	0,63	1	
Tegnsetting	0,56	0,53	0,55	0,62	0,63	1

Prinsipal komponentanalysen viser en sterkere tendens til en-faktorstruktur enn på 5. trinn (se tabell 4.5). Denne forskjellen har vi også sett ved tidligere gjennomganger, som for eksempel våren 2015. Videre ser vi samme mønster som tidligere med at vurderingsområdene kommunikasjon, innhold, tekstoppbygging og språkbruk lader sterkt opp mot den ene faktoren som i dette tilfellet forklarer 69 % av total varians. I den andre faktoren (kun 0,67 eigenvalue) lader de to siste vurderingsområdene sterkt. Vi har likevel inkludert to faktorer i rotasjonen på samme måte som på 5.trinn.

Tabell 4.5 Prinsipal komponentanalyse av vurderingsområdene på 8.trinn

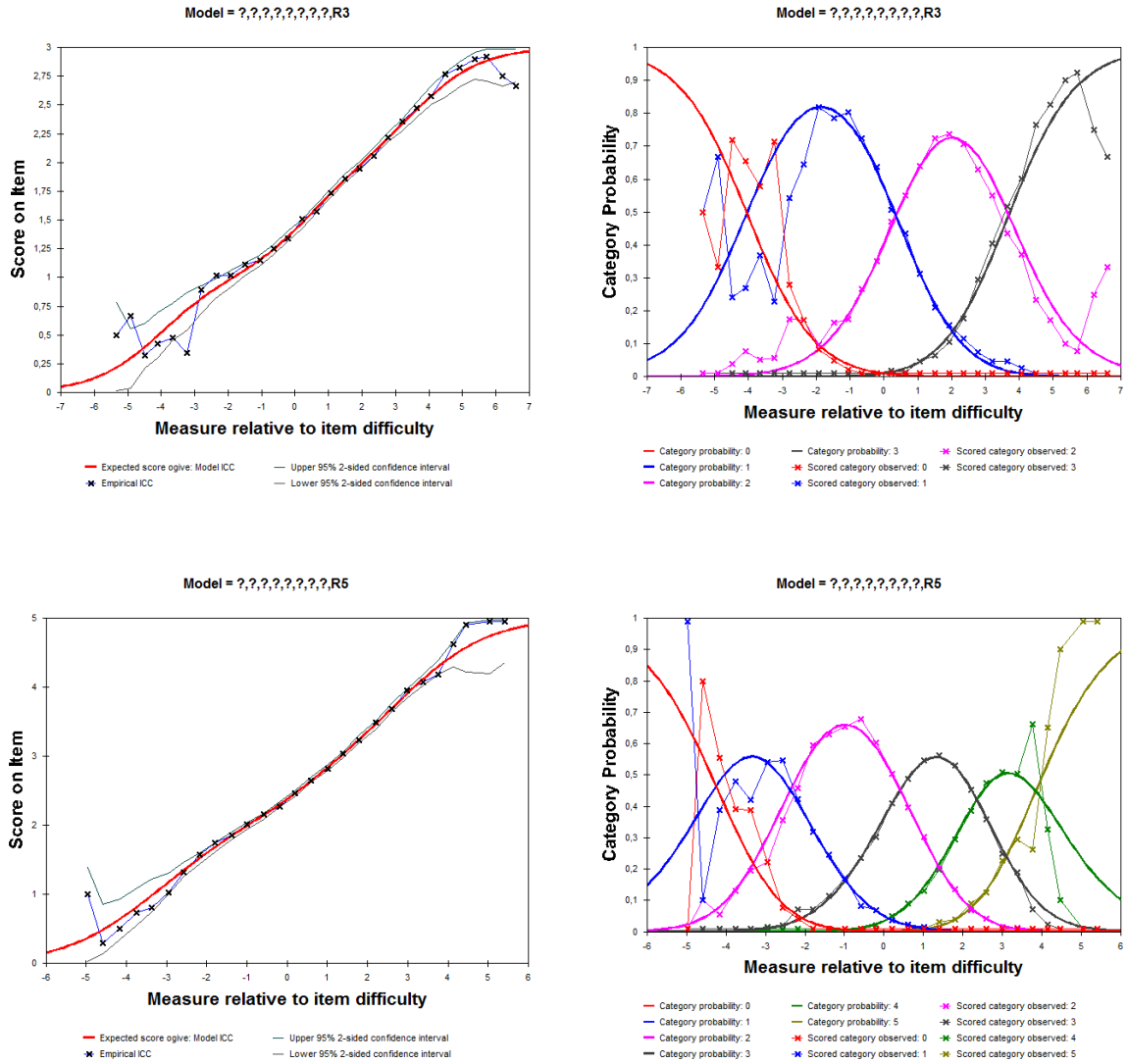
	Faktor 1	**Faktor 2**
Eigenvalue	4,158 (69%)	0,67 (11 %)
Kommunikasjon	0,887	0,038
Innhold	0,967	-0,088
Tekstoppbygging	0,888	0,009
Språkbruk	0,588	0,365
Rettskriving	-0,017	0,917
Tegnsetting	0,030	0,869

Konklusjonen av analysene av faktorstrukturen er derfor at faktorstrukturen er tilnærmet som tidligere og vi ser en relativt klar en-faktorstruktur hvor fire vurderingsområder lader sterkt.

4.2 Fit, reliabilitet, och skalstruktur

Den grafiska inspektionen indikerar att data passar modellen väl, både på 5 och på 8. trinn. I figur 4.1 nedan återges en matris med den aktuella grafiken.

Figur 4.1 Grafisk inspektion av relationen data–modell



Notera. 5. trinn: överst i figuren, 8. trinn: underst i figuren. Grafik-rubrikerna avser Facets-specifikationer, nämligen antalet facets och högsta tillåtna poäng.

I figuren kan vi utläsa följande. För 5. trinn, som representeras av de två översta figurerna passar data modellen bra, de modellförväntade resultaten stämmer i de flesta fall överens med de empiriska. Det samma gäller för 8. trinn. För de lägsta och högsta kategorierna (dvs. 0 och

3 respektive 5) finns viss diskrepans mellan modellförväntningar och empiriska observationer. Detta beror dels på sättet på vilket logit estimeras (den matematiska modellen implementerad i *Facets*), dels på att antalet observationer där är få.

4.2.1 Elever, uppgifter, bedömningsområden och provtillfälle

I tabell 4.6 vi ser på övriga index och får indikationer på huruvida mätningen verkar meningsfull. Vi noterar att mätverktygen möjliggjort signifikant diskriminering mellan nivåer av skrivförmåga på båda trinnen (jfr *Q*, *H*- och *R*-index för elever).

För elever kan vi också notera att andelen signifikanta infit och outfit-värden som indikerar ”overfit” (dvs. som understiger 0,75; $t > 2,0$) är något lägre än andelen signifikanta infit- och outfit-värden som indikerar ”underfit” (dvs. som överstiger 1,30; $t > 2,0$). Underfit indikerar att elever har fått ”oväntade” resultat, vilket i sin tur sannolikt kommer av bristande samstämmighet bland bedömarna. Överlag är dock inte andelen over- och underfit oroväckande hög.[6]

Tabell 4.6 Kvalitetsindikatorer

		RMSE	True S.D.	Q	H	R	Infit Low	Infit High	Outfit Low	Outfit High	Enig %	SR-ROR	ICC
5 t	Elev.	0,39	1,60	5164,5**	5,73	,94	5,9	7,8	6,3	8,1	-	-	-
	Uppg.	0,03	0,42	200,2**	19,15	1,00	0,0	0,0	0,0	0,0	-	-	-
	Skal.	0,05	0,24	115,7**	6,63	,96	0,0	0,0	0,0	0,0	-	-	-
	Bed.	0,12	0,53	660,6**	6,18	,95	5,9	2,9	20,6	5,9	57,9	,25	,65
8 t	Elev.	0,29	1,44	11000,3**	6,94	,96	9,1	9,7	8,5	10,3	-	-	-
	Uppg.	0,02	0,16	67,9**	11,24	,99	0,0	0,0	0,0	0,0	-	-	-
	Skal.	0,03	0,24	243,7**	9,57	,98	0,0	0,0	0,0	0,0	-	-	-
	Bed.	0,09	0,58	1728,3**	8,92	,98	19,0	16,7	19,0	11,9	45,8	,27	,67

Notera. Elev. = Elever; Uppg. = Uppgifter; Skal. = Skalor/bedömningsområden; Bed. = Bedömare; RMSE = Root Mean Square Error; True S.D. = sann varians; Q = homogenitetsindex; H = separationsindex; R = reliabilitetsindex; Infit Low = % statistiskt signifikant infit < 0,75; Infit High = % statistiskt signifikant infit > 1,30; Outfit Low = % statistiskt signifikant infit < 0,75; Outfit High = % statistiskt signifikant infit > 1,30; Enig % = procent fullständig samstämmighet; SR-ROR = Single Rater-Rest of Raters; ICC = Intraclass Correlation Coefficient average measure. ** p = < ,01.

Går vi vidare till uppgifterna ser vi att vi även här har att göra med signifikanta skillnader på båda trinnen. Detsamma gäller skalorna på bägge trinnen. Det sistnämnda resultatet är i linje med resultat som rapporterats i tidigare tekniska rapporter, men det första är en nyhet.

[6] Infit- och outfit-värden, som indikerar data-modell-relation på elementnivå, antar vanligen ett genomsnitt på 1,0 (som är det förväntade värdet) och rapporteras därför här som andelar.

Eftersom U2015 är första provet som inkluderat två uppgifter till samma elever kan vi nu med säkerhet slå fast att uppgifterna är olika svåra.

Avslutningsvis, och relaterat till frågan om uppgifters svårighet skall vi se närmare på om tidpunkten för genomförande av provet inverkar på resultatet. i MFRM-analysen ingår tidpunkt som ett dummy-facet, vilket innebär att något logit-estimat inte beräknas. Därför får vi använda råpoängen. I tabell 4.7 se vi skillnader, uttryckta i genomsnittligt råpoäng mellan provtillfälle 1 och 2.

Tabell 4.7 Skillnader mellan tillfälle 1 och tillfälle 2

	Tillfälle 1	**Tillfälle 2**
5. trinn	1,81	1,67
8. trinn	2,75	2,72

Som vi kan se finns det en marginell skillnad på 5. trinn, men ingen nämnvärd skillnad på 8. trinn. Detta kan förklaras med att elever på 5. trinn är mindre vana vid formella provsituationer. Tack vare designen har inte skillnaderna mellan provtillfällen någon betydelse när vi skall tolka skillnader i uppgifters svårighetsgrad.

4.2.2 Bedömare

Vad bedömarna beträffar kan vi konstatera att de inte är utbytbara med varandra. Vi ser att det på bägge trinnen finns signifikanta och trovärdiga skillnader i stränghet (vilket indikeras av *Q*-, *H*- och *R*-index). På 5. trinn är antalet statistiskt distinkta klasser av stränghet 6,2 och på 8. trinn 8,9. Vi ser, när vi granskar fit-statistiken, att tendensen till overfit är vanligare än tendensen till underfit. Det betyder att bedömarna, på båda trinnen, i högre utsträckning uppvisar liten variation. Detta kommer typiskt av *halo*, vilket innebär att omdömet på ett område färgar av sig på omdömet på andra områden, eller *centraltendens*. Andelen underfit på 8. trinn är högre än andelen på 5. trinn, men inte oroväckande hög.

Den procentuella samstämmigheten är högre på 5. trinn än på 8. trinn, men där är också skalorna kortare. SR-ROR-värdena indikerar att panelen på bägge trinnen befinner sig precis i underkant för vad som är gynnsamt när bedömardata skall passas till MFRM-modellen. Tack vare att bedömarna är tillräckligt konsistenta (jfr fit-statistik och reliabiliteten för H-index) får detta dock ingen avgörande betydelse.

Det sista måttet, ICC, är samstämmigt med övriga mått i det att det ligger under vad som är eftersträvansvärt vid traditionell behandling av rådata som bygger på bedömningar. Två förhållanden är emellertid viktiga att ta hänsyn till vid tolkning av ICC. För det första bygger det inte på totalpoäng, vilket gör det mindre jämförbart med *Facets*-indexen, som ju tar hänsyn till all tillgänglig information. För det andra, vilket framkommer i tabell 4.8, kan vi se en viss variation mellan bedömningsområden.

Tabell 4.8 ICC Average för olika bedömningsområden

	V1	V2	V3	V4	V5	V6	Gjennomsn.
5. trinn	,60	,64	,54	,64	,75	,73	,65
8. trinn	,64	,59	,66	,65	,76	,69	,67

Notera. V1 = Kommunikasjon; V2 = Innhold; V3 = Tekstoppbygging; V4 = Språkbruk; V5 = Rettskriving; V5 = Tegnsetting.

Generellt finner vi högst ICC-värden för de skalor som mäter kodningskompetenserna, det vill säga V5 och V6. Dessa områden har tidigare uppvisat låg grad av bedömarsamstämmighet. På 5. trinns är det särskilt V1 och V3 som drar ner ICC-värdena, medan det framstår som att V2 är mest problematisk på 8. trinn. Dessa områden kan därför vara i behov av särskild uppmärksamhet när nya vägledningar för Skriveprøven 2016 skrivs. Skrivesenteret kommer att återkomma till frågan om bedömarpanelen i en senare rapport.

4.2.3 Skalstruktur

Vi skall också se närmare på den generaliserade skalan. Till grund för U2015 ligger förvisso strikt analytisk bedömning, men i den MRFM-analys vi har genomfört generaliseras skalorna. Det finns ett antal kvalitetsindikatorer (Eckes, 2015, s. 117ff): För att skalan skall anses fungera bra måste varje skalsteg vara observerat minst tio gånger. Varje skalsteg måste dessutom vara det mest sannolika vid någon given förmågenivå. Vi har tidigare nämnt Rasch-Andrich Thresholds som ett viktigt mått. Detta bör avancera med minst 1,4 logits, men inte med mer än 5,0 logits. Dessutom är det av betydelse att fit-statistiken indikerar att den modellförväntade förmågan relaterad till ett visst skalsteg motsvarar den observerade. För att så skall vara fallet bör inte fit-statistiken överstiga 2,0. I tabell 4.9 och tabell 4.10 återges resultaten för 5. trinn respektive 8. trinn.

Tabell 4.9 Skalstruktur och kvalitetsindikatorer, 5. trinn

Kat.	Antal	Andel (%)	Logit m.	Logit exp	Outfit	R-A T	S.E.
0	167	2	-2,86	2,41	0,8	-	-
1	2 943	34	-0,01	-0,10	1,1	-4,05	0,09
2	4 375	51	1,48	1,55	1,1	0,36	0,03
3	1 065	12	3,10	3,00	0,9	3,69	0,04

Notera. Kat = kategori; Logit m = genomsnittligt logitvärde för kategori; Logit exp = förväntat logitvärde för kategori; R-A T = Rasch-Andrich Threshold; S.E. = mätfelet för R-A T.

Tabell 4.10 Skalstruktur och kvalitetsindikatorer, 8. trinn

Kat.	Antal	Andel (%)	Logit m.	Logit exp	Outfit	R-A T	S.E.
0	94	1	-2,90	2,67	0,9	-	-
1	831	8	-1,43	1,47	1,0	-4,25	0,11
2	3 838	35	-0,19	-0,20	1,0	-2,37	0,04
3	3 918	35	1,03	1,05	1,0	0,41	0,02
4	1 851	17	2,18	2,13	1,0	2,36	0,03
5	508	5	2,87	2,93	1,0	3,85	0,05

Notera. Kat = kategori; Logit m = genomsnittligt logitvärde för kategori; Logit exp = förväntat logitvärde för kategori; R-A T = Rasch-Andrich Threshold; S.E. = mätfelet för R-A T.

Vi ser att för bägge trinnen gäller att alla kategorier är observerade vid minst 10 tillfällen. Vi kan också se konturerna av en unimodal fördelning, när vi granskar andelen för respektive, oberoende av trinn. Ser vi därefter på observerade och förväntade logitvärden för respektive kategori, tillsammans med outfit-värden, kan vi notera att observationerna ligger mycket nära det förväntade värdet på bägge trinnen. Det är också så att högre kategorivärden är förbundna med högre logitvärden. Vi ser också att Rasch-Andrich Thresholds avancerar enligt nyssnämnda kriterier, även här för bägge trinnen.

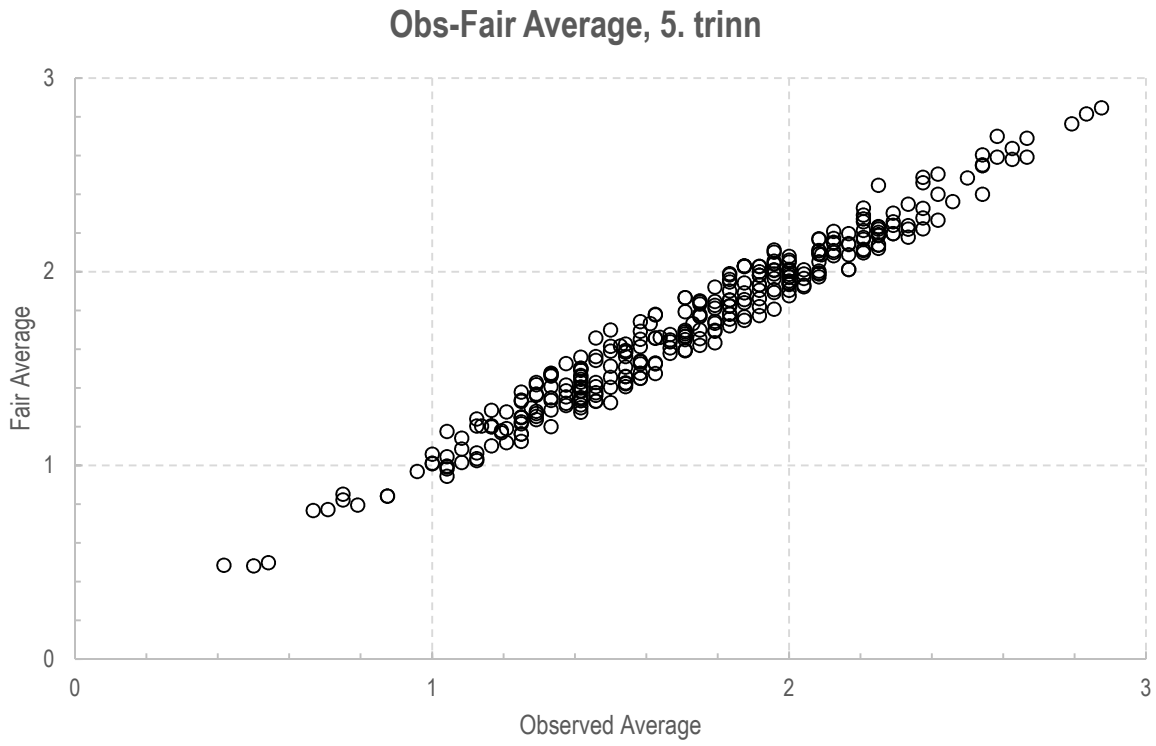
4.3 Sammanfattning – från observerad till rättvis poäng

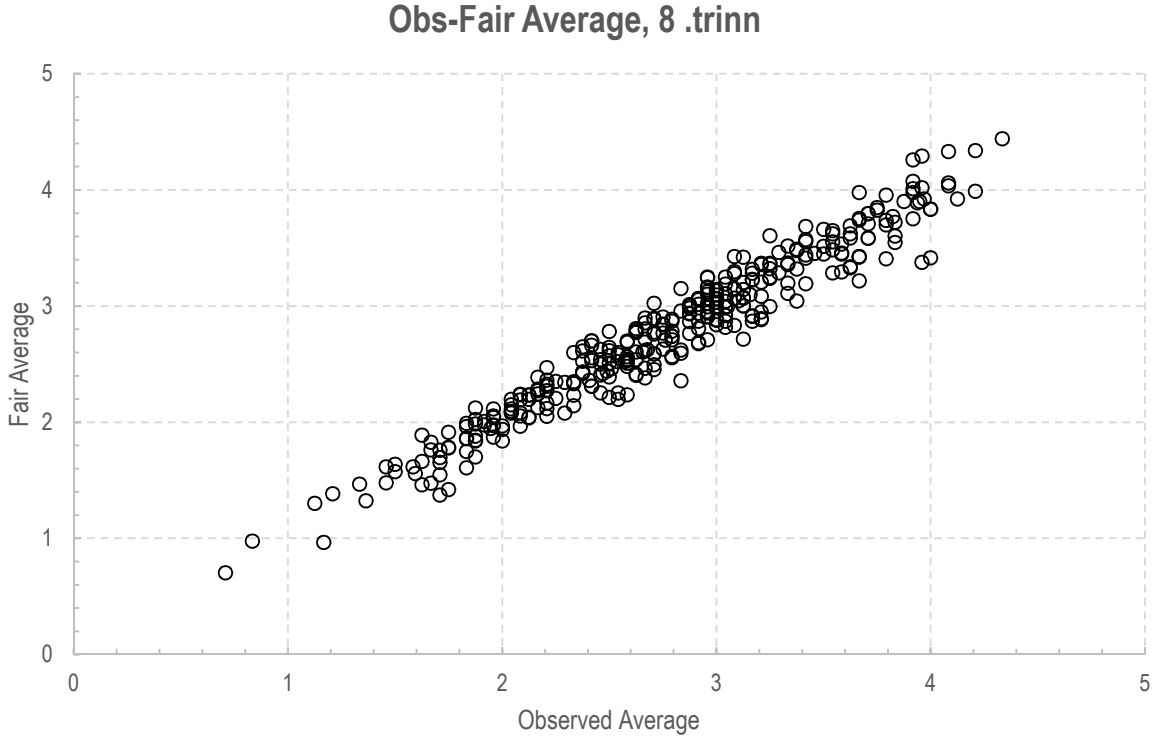
I detta kapitel har vi studerat de psykometriska kvaliteterna förbundna med U2015. Faktor-analyser och estimat av ”global fit” ger starka indikationer för att data passar modellen. Övriga kvalitetsindikatorer avslöjar att olika facets bidrar till variansen, inte minst bedömarna. Resultaten visar dock att rådata har kunnat passas till modellen på ett trovärdigt sätt, vilket innebär att det finns all anledning att fästa tilltro till de skalpoäng som presenteras i nästa kapitel. Sålunda har MFRM-modellen varit ändamålsenlig för att kunna ta fram en nationell nivå.

För att illustrera vad MFRM-analysen innebär i termer av konvertering av råpoäng visar vi avslutningsvis ett diagram för respektive trinn som illustrerar relationen mellan observerade råpoäng och poäng på fair average-skalan. Graferna visar att de flesta elever får ett något

justerat resultat när vi tar hänsyn till bedömares stränghet, men att justeringarna i allmänhet när tämligen små.

Figur 4.2 Relationen observerad och fair average för 5. trinn och 8. trinn





5 Nationell nivå

I detta kapitel presenteras den nationella nivån på gruppnivå och sub-gruppnivå. Vi kommer att presentera s.k. Wright-maps, som är grafiska sammanställningar av MFRM-analysen och enkla medelvärdesanalyser för respektive trinn. Vi skall också säga något om hur logit-poängen har skalats om till nationell nivå-skala.

På 5. trinn var medelvärdet för elever 1,04 logits (fair average: 1,73) och på 8. trinn var medelvärdet 0,75 (fair average: 2,77). För att detta skall motsvara 50,0 har vi enlighet med specifikationer i Linacre (2013) utfört en linjär transformering. Detta har resulterat i följande ekvationer för 5. trinn respektive 8. trinn:

$$x_{\text{nationell\_nivå\_5}} = y_{\text{5\_trinn}} * 5{,}23 + 44{,}58$$

$$x_{\text{nationell\_nivå\_8}} = y_{\text{8\_trinn}} * 6{,}17 + 45{,}35$$

där

x = poäng på nationell nivå

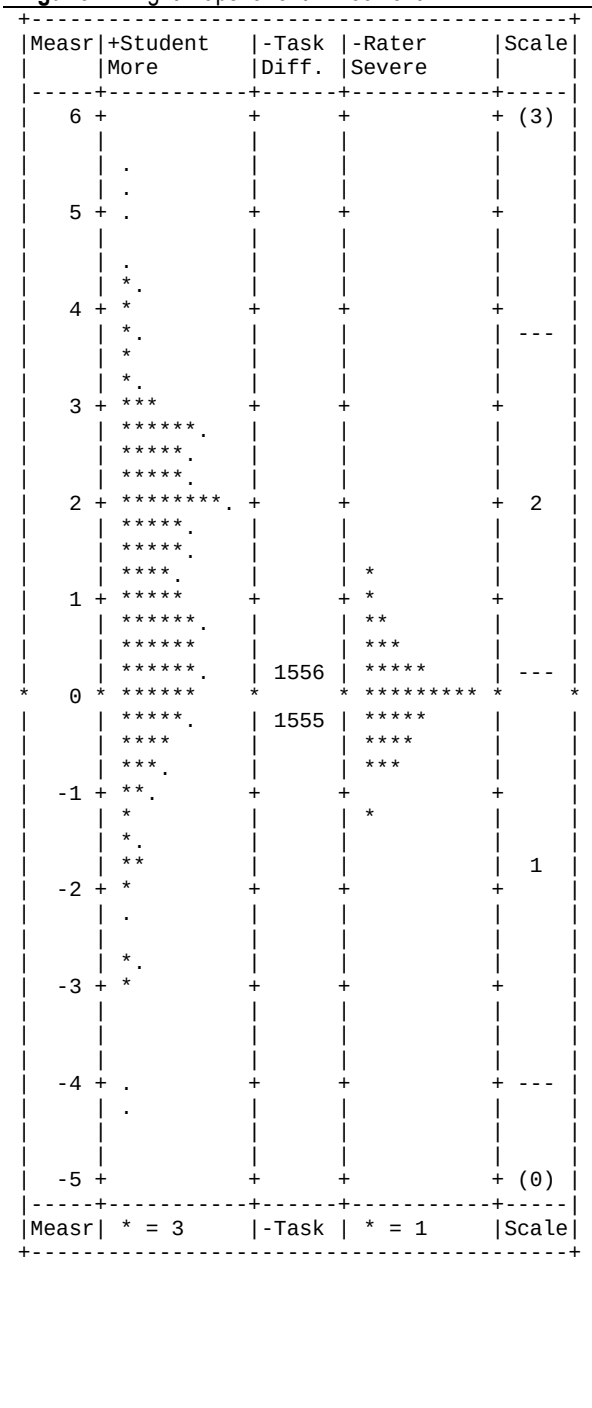
y = logit-poäng[7]

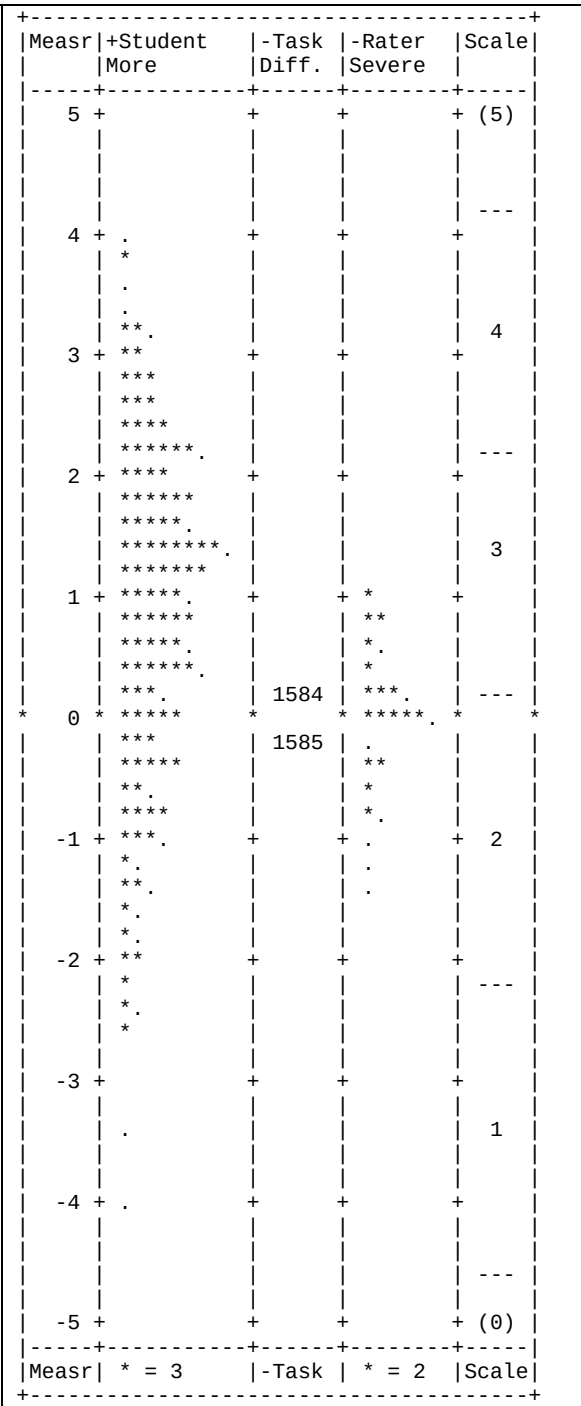
Den tekniska degression som är nödvändig för att med någorlunda precision förklara varför ekvationerna skiljer sig åt från 5. trinn till 8. trinn avstår vi från av utrymmesskäl. I stället konstaterar vi att denna enkla transformering gör att elever på 5. trinn och på 8. trinn får samma medelvärde. Läsaren bör dock hålla i minnet att resultaten mellan trinnen inte är jämförbara, eftersom olika skalor och olika uppgifter har använts.

I figur 5.1 nedan presenteras Wright-maps för repsektive trinn. Från vänster till höger ser vi logitskalan, elevresultaten, uppgiftssårigheten, bedömarsträngheten och den ursprunliga skalan. Notera att för elever och bedömaren kan kartan avläsas som ett histogram.

[7] Den uppmärksamma läsaren noterar att ekvationen inte ger fullt förväntat resultat. I transformeringen användes därför fler decimaler. På 5. trinn, således: y5 * 5,22793810121288 + 44,5760508155583, och på 8. trinn: y8 * 6,17344922955354 + 45,350176560648.

Figur 5.1 Wright-maps för 5. trinn och 8. trinn





Notera. Kartan för 5. trinn till vänster och kartan för 8. trinn till höger. Ju högre upp en elev befinner sig, desto högre resultat på utvalgsprøven. För uppgifter och bedömare gäller att ju högre upp, desto svårare, respektive strängare. Notera att logit-skalan går från cirka -5,0 till 5,0. Detta är bara av konventionella skäl. Ingen elev har allstå ”negativ förmåga”.

Vi ser att det finns en betydande spridning i elevgruppen, liksom vi redan konstaterat att det finns en icke obetydlig spridning bland bedömarna. Vi ser också att båda de uppgifter som efterfrågar skrivhandling att föreställa sig är enklare än de andra uppgifterna. I tabellerna 5.1 och 5.2 skall vi se närmare på resultaten på grupp- och subgruppnivå.

Tabell 5.1 Grupp- och sub-gruppnivå, 5. trinn

	n	M	S.E.M.	S.D.	*t*	*df*	*p*	*ES*
Alla elever	320	50,0	2,05	8,6				
Flickor	166	53,3	2,04	7,4				
Pojkar	154	46,5	2,06	8,4				
Diff flickor-pojkar		6,8			7,65	318	,00	0,86
BM	264	49,4	2,06	8,7				
NN	56	52,7	2,03	7,4				
Diff BM-NN		-3,3			-2,65	318	,00	-0,39
L1	295	50,3	2,05	8,5				
L2	25	46,8	2,05	9,2				
Diff L1-L2		3,5			1,97	318	,05	0,41

Notera. M = medelvärde, S.E.M. = standard error of measurement (MFRM-baserat); S.D. = standardavvikelse; t = t-värde; df = degrees of freedom; ES = effect size (Cohens's *d*, ojusterad).

På 5. trinn föreligger signifikanta skillnader mellan varje sub-grupp. Den största skillnaden återfinns mellan flickor och pojkar. Flickorna presterar nästan en standardavvikelse bättre än pojkarna. Annorlunda uttryckt motsvarar skillnaden en effektstorlek om d = 0,86, vilket kan anses vara substantiellt (Cohen, 1988). (Se appendix för tolkning av d-värden). Skillnaderna mellan övriga grupper än mindre med effektstorlekar på omkring d = 0,40.

Tabell 5.2 Grupp- och sub-gruppnivå, 8. trinn

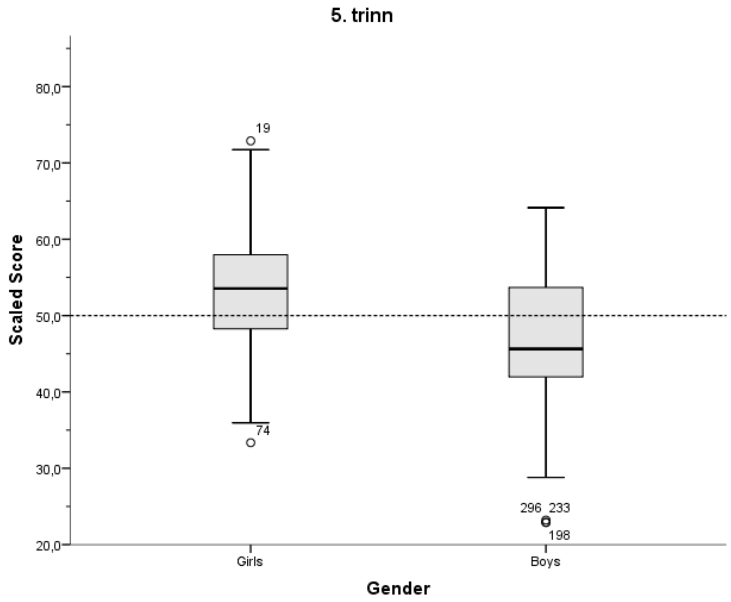
	n	M	S.E.M.	S.D.	t	df	p	ES
Alla elever	351	50,0	1,78	9,1				
Flickor	196	52,3	1,75	8,1				
Pojkar	155	47,1	1,80	9,4				
Diff flickor-pojkar		5,3			5,6	349	,00	0,60
BM	247	50,2	1,78	9,4				
NN	104	49,6	1,77	8,2				
Diff BM-NN		0,6			0,5	349	,59	0,06
L1	317	50,5	1,77	8,9				
L2	34	45,2	1,84	9,3				
Diff L1-L2		5,3			3,3	349	,00	0,59

Notera. M = medelvärde, S.E.M. = standard error of measurement (MFRM-baserat); S.D. = standardavvikelse; t = t-värde; df = degrees of freedom; ES = effect size (Cohens's *d*, ojusterad).

På 8. trinn är skillnaderna mellan grupper överlag något mindre. Vi ser dock att flickor och pojkar får signifikant olika resultat även här, med en effektstorlek på d = 0,60, något kan betraktas som moderat. Målform innebär ingen skillnad på 8. trinn. Däremot är skillnaden mellan elever med norska som L1 och norska som L2 är något högre på 8. trinn (effektstorlek d = 0,59).

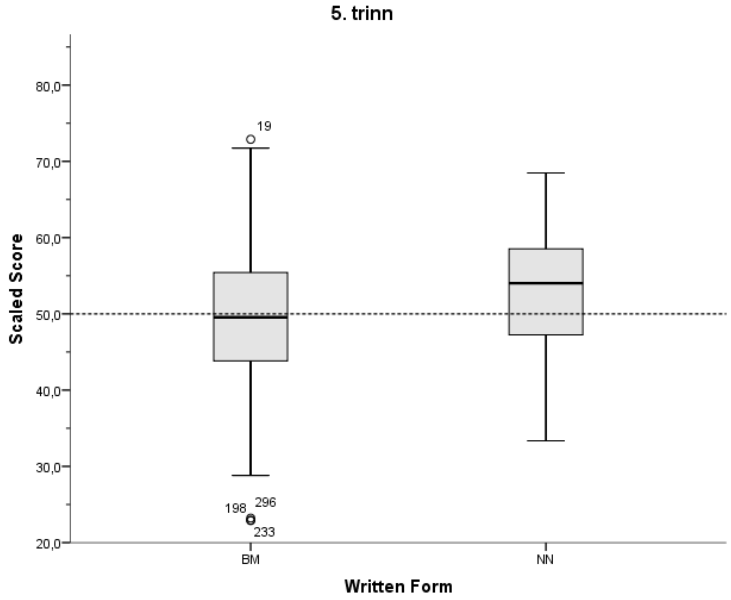
I figurerna 5.2–5.7 åskådliggörs skillnaderna mellan grupperna på 5. trinn och 8. trinn grafiskt.

Figur 5.2 Skillnader mellan flickor och pojkar, 5. trinn



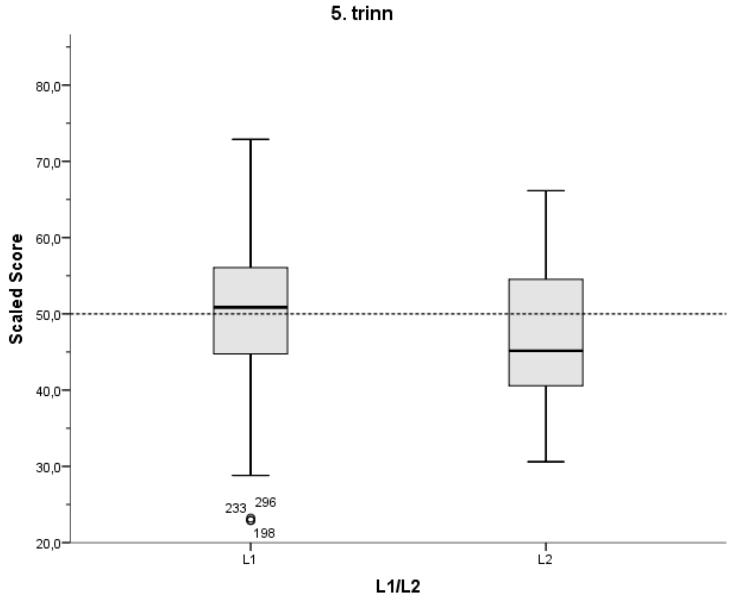
Notera. Medianvärde: lådans mitt. Undre och övre kvartil: lådans kanter. ”Morrhåren” leder till min-/maxvärde. Ringar: ”uteliggare” (outliers). Streckad linje: medelvärdet (50,0).

Figur 5.3 Skillnader mellan bokmåls- och nynorskelever, 5. trinn



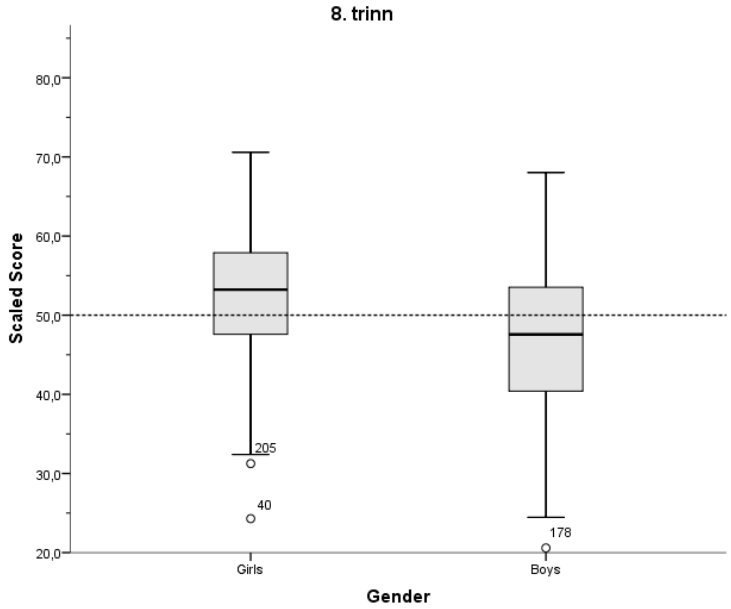
Notera. Medianvärde: lådans mitt. Undre och övre kvartil: lådans kanter. ”Morrhåren” leder till min-/maxvärde. Ringar: ”uteliggare” (outliers). Streckad linje: medelvärdet (50,0).

Figur 5.4 Skillnader mellan elever med norska som L1 och L2, 5. trinn



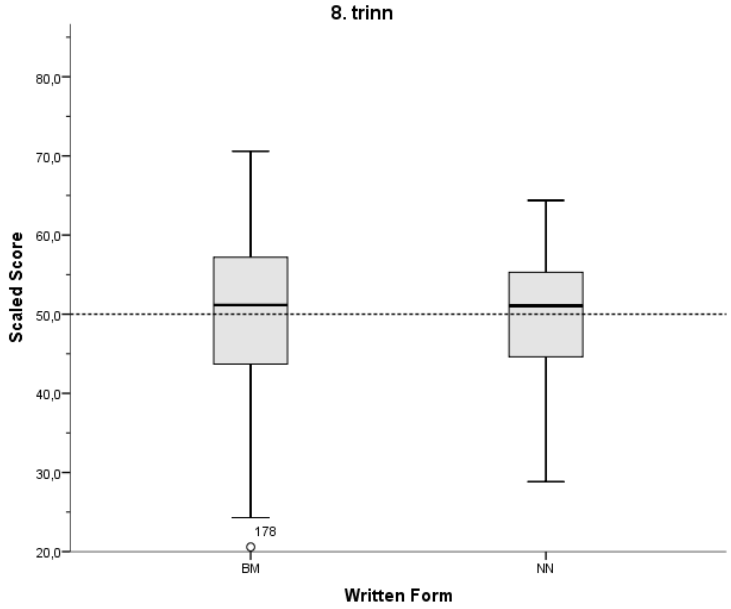
Notera. Medianvärde: lådans mitt. Undre och övre kvartil: lådans kanter. ”Morrhåren” leder till min-/maxvärde. Ringar: ”uteliggare” (outliers). Streckad linje: medelvärdet (50,0).

Figur 5.5 Skillnader mellan flickor och pojkar, 8. trinn



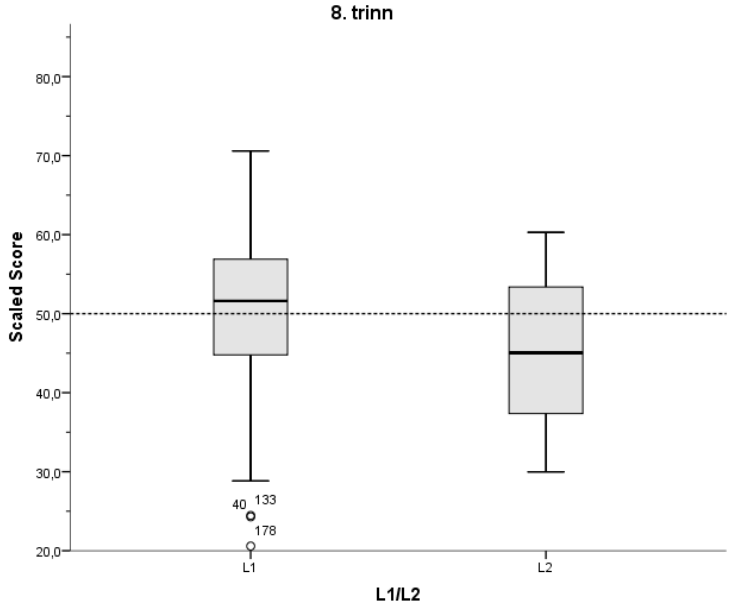
Notera. Medianvärde: lådans mitt. Undre och övre kvartil: lådans kanter. ”Morrhåren” leder till min-/maxvärde. Ringar: ”uteliggare” (outliers). Streckad linje: medelvärdet (50,0).

Figur 5.6 Skillnader mellan bokmåls- och nynorskelever, 8. trinn



Notera. Medianverde: lådans mitt. Undre och övre kvartil: lådans kanter. "Morrhåren" leder till min-/maxvärde. Ringar: "uteliggare" (outliers). Streckad linje: medelvärdet (50,0).

Figur 5.7 Skillnader mellan elever med norska som L1 och L2, 8. trinn



Notera. Medianverde: lådans mitt. Undre och övre kvartil: lådans kanter. "Morrhåren" leder till min-/maxvärde. Ringar: "uteliggare" (outliers). Streckad linje: medelvärdet (50,0).

Graferna visar både skillnader mellan grupper och också spridning inom grupper. Att grupperna presterar så olikt på aggregerad nivå har gjort det relevant att studera skillnaderna närmare.

För att göra det har vi genomfört MFRM-analyser med fokus på bedömningsområden och på uppgifter.[8] För att underlätta tolkningen av resultaten presenteras de på *fair average*-skalan.

Tabell 5.3 Detaljerad sub-grupp-analys, 5. trinn

		M_1	$S.D._1$	M_2	$S.D._2$	*t*	*df*	*p*	*ES*
Funk.	Flickor-Pojkar	1,92	0,40	1,55	0,43	8,1	318	,00	0,89
	BM-NN	1,71	0,46	1,89	0,43	-2,6	318	,01	-0,40
	L1/L2	1,76	0,45	1,52	0,46	2,6	318	,01	0,53
Kod.	Flickor-Pojkar	1,83	0,45	1,54	0,49	5,6	318	,00	0,62
	BM-NN	1,66	0,50	1,82	0,41	-2,3	318	,02	-0,33
	L1/L2	1,69	0,49	1,63	0,56	0,6	318	,58	0,12
U1555	Flickor-Pojkar	1,96	0,48	1,63	0,46	6,3	317	,00	0,70
	BM-NN	1,77	0,49	1,95	0,48	-2,5	317	,01	-0,37
	L1/L2	1,82	0,49	1,59	0,48	2,2	317	,03	0,47
U1556	Flickor-Pojkar	1,83	0,42	1,45	0,51	7,1	318	,00	0,82
	BM-NN	1,61	0,51	1,81	0,42	-2,7	318	,01	-0,40
	L1/L2	1,66	0,50	1,52	0,55	1,3	318	,20	0,28

Notera. Funk = funktionskompetenser, kod. = kodningskompetenser; M_1 = Medelvärde grupp till vänster; $S.D._1$ = standardavvikelse grupp till vänster; M_2 = Medelvärde grupp till höger; $S.D._2$ = standardavvikelse grupp till höger; t = t-värde; df = degrees of freedom; EF = effect size (Cohen's *d*, ojusterad).

Även med den förfinade analysen framgår att de största skillnaderna på 5. trinn återfinns mellan flickor och pojkar. Störst är den på funktionskompetenser, d = 0,89, och minst på kodningskompetenser, d = 0,62. Den minsta skillnaden mellan flickor och pojkar är dock större än någon annan skillnad mellan andra grupper.

Som syns i tabellen finns några icke-signifikanta skillnader mellan elever som har norska som L1 och elever som har norska som L2. En möjlig förklaring vid sidan av att skillnaderna de facto är små, är att gruppen L2-skribenter är för liten för att den statistiska styrkan skall bli tillräcklig.

Värt att notera är också att skillnaden mellan elever som skriver på bokmål och på nynorsk är störst vad gäller funktionskompetanser. Vidare undersökningar kan fokusera på om detta betyder att målformen i sig inte är direkt utslagsgivande.

Vänder vi så blicken mot 8. trinn kan vi i tabell 5.4 se att den förfinade analysen ger oss viss ny insikt. Skillnader mellan flickor och pojkar är förvisso fortsatt den som dominerar, men

[8] Dessa analyser har genomförts *utan* ankarvärden från huvudanalysen. Precisionen blir då något lägre, men resultaten kan användas diagnostiskt i det vidare arbetet. Alla specifikations- och resultatfiler lämnas tillhandahålls av huvudförfattaren.

den finmaskiga analysen visar att det föreligger *en* signifikant skillnad mellan elever med bokmål och elever med nynorsk. Till skillnad från 5. trinn är det de förra presterar bättre, denna gång på skalor som mäter kodningskompetenser. Skillnaden är dock liten och effektstorleken likaså, d = 0,26. Även här vore fortsatta undersökningar av stort intresse.

Till skillnad från 5. trinn är skillnaderna mellan elever med norska som L1 och L2 signifikant på alla undersökta variabler. Största skillnaderna ser vi på skalor som mäter funktionskompetenser och på U1585. Det förra resultatet kan möjligen förklaras av att skalorna som mäter funktionskompetens fångar upp typisk andraspråksproblematik, men det senare resultatet är mer svårförklarat. Vidare analyser får svara på detta.

Tabell 5.4 Detaljerad sub-grupp-analys, 8. trinn

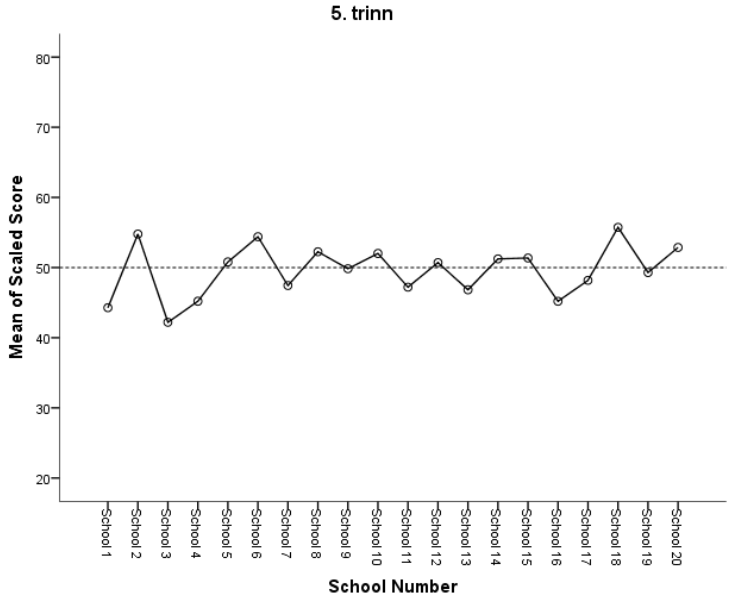
		M_1	S.D.$_1$	M_2	S.D.$_2$	*t*	*df*	*p*	*ES*
Funk.	Flickor-Pojkar	2,96	0,66	2,56	0,71	5,45	349	,00	0,59
	BM-NN	2,78	0,75	2,79	0,63	-0,13	349	,90	-0,02
	L1/L2	2,83	0,70	2,40	0,66	3,43	349	,00	0,62
Kod.	Flickor-Pojkar	2,90	0,65	2,54	0,71	4,91	349	,00	0,53
	BM-NN	2,80	0,72	2,62	0,64	2,21	349	,03	0,26
	L1/L2	2,77	0,69	2,48	0,74	2,36	349	,02	0,43
U1584	Flickor-Pojkar	2,86	0,67	2,55	0,77	4,04	349	,00	0,43
	BM-NN	2,72	0,76	2,73	0,68	-0,16	349	,87	-0,02
	L1/L2	2,75	0,74	2,49	0,67	2,00	349	,05	0,36
U1585	Flickor-Pojkar	3,03	0,75	2,59	0,76	5,46	349	,00	0,59
	BM-NN	2,85	0,82	2,79	0,70	0,68	349	,50	0,08
	L1/L2	2,88	0,77	2,39	0,79	3,56	349	,00	0,64

Notera. Funk = funktionskompetenser, kod. = kodningskompetenser; M_1 = Medelvärde grupp till vänster; S.D.$_1$ = standardavvikelse grupp till vänster; M_2 = Medelvärde grupp till höger; S.D.$_2$ = standardavvikelse grupp till höger; t = t-värde; df = degrees of freedom; EF = effect size (Cohen's *d*, ojusterad).

Som en sista analys har vi också jämfört resultaten på skolnivå. Fynd från stora internationella undersökningar brukar ofta peka i riktning mot att variansen är större inom än mellan skolor. I vårt fall rör vi oss på klassnivå så vi förväntar ett annat resultat.

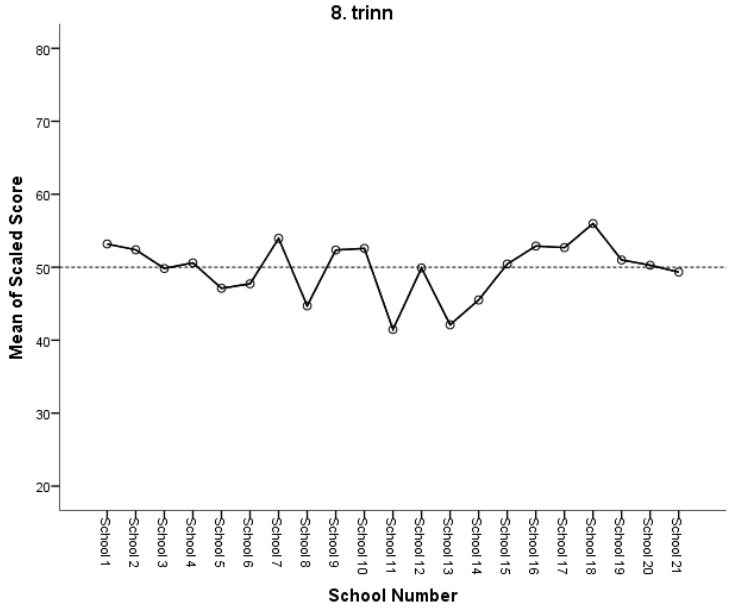
I figur 5.5 och 5.6 återfinns en medelvärdesgrafer, som illustrerar skillnaderna mellan skolor på respektive trinn. Grafen visar att det finns betydande skillnader mellan skolor. Detta bekräftas av envägs-ANOVA. För 5. trinn är skillnaderna mellan skolor signifikant, F(19, 300) = 2,9, p = < ,00, liksom de är för skolor på 8. trinn, F(20, 330) = 3,1, p = < ,00. Försiktigt uttryckt kan vi därför konstatera att flera variabler bidrar till variation. Vid sidan av kön och L1/L2 tycks också skola spela en avgörande betydelse för resultatet.

Figur 5.5 Resultat på skolnivå, 5. trinn



Notera. Streckad linje: medelvärdet (50,0).

Figur 5.6 Resultat på skolnivå, 8. trinn



Notera. Streckad linje: medelvärdet (50,0).

Avslutning

Syftet med denna tekniska rapport har varit att beskriva arbetet med att mäta elevers skrivförmåga inom ramen för utvalgsprøven och att redovisa den nationella nivån på olika nivåer.

Resultaten visar att MFRM-analysen på ett trovärdigt sätt kunnat modellera rådata för att ta fram nationell nivå. Resultaten visar också att det utan en MFRM-analys skulle vara tveksamt att dra några säkra slutsatser om elevers skrivförmåga. Detta beror på att bedömargruppen står för en inte betydelselös variation. Skrivesenteret kommer att författa en detaljerad rapport som avhandlar bedömargruppen. I övrigt är resultaten från den psykometriska kvalitetsundersökningen viktiga för det fortsatta arbetet med utveckling av Skriveprøven och av arbetsmaterial för bedömarpanelen.

När det gäller den nationella nivån kan konstateras att årets utvalsgprøve kunnat visa att uppgifterna är olika svåra. De uppgifter som bygger på skrivhandlingen *att föreställa sig* är enklare än andra uppgifter.

Resultaten visar också att det finns betydande skillnader mellan olika grupper. Störst är skillnaden mellan flickor och pojkar, oavsett trinn. Skillnaden mellan elever med bokmål och elever med nynorsk är långt ifrån entydig. På 5. trinn är den signifikant oavsett om vi undersöker kompetenstyp eller uppgifter. På 8. trinn ser skillnaderna till stor del ut att ha försvunnit, men på skalor som mäter funktionskompetenser finns de kvar. Dock i förändrad riktning, så att det är elever med bokmål som presterar bättre.

Elever med norska som L1 presterar i allmänhet bättre än elever med norska som L2. Skillnaderna är tydligare på 8. trinn än på 5. trinn. Detta kan ha att göra med svag statistisk styrka, men – förstås – också att skillnaderna faktiskt inte är reella.

Slutligen skall sägas något om skolorna. Vi har kunnat visa en betydande skillnad mellan skolor och variansen är större mellan än inom skolor. Detta beror sannolikt på en mängd faktorer, så som att urvalsstorlekarna skiljer sig åt och att elever är samplade från klasser och därmed inte representerar hela skolor. Oavsett vad skillnaderna beror av är de viktiga att

uppmärksamma. Resultaten både för 5. trinn och 8. trinn visar att elever från olika skolor har tämligen olika förutsättningar för att lösa uppgifterna i utvalgsprøven.

Denna rapport har inneburit meningsfulla analyser för det fortsatta arbetet med provutveckling. Förhoppningsvis är den också värdefull för skolsektorn. Resultaten visar att det finns behov av utveckla en skrivundervisning som passar alla elever.

Litteratur

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Modern English Publications/British Council/Macmillan.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1301–1322). Chichester, West Sussex: Wiley-Blackwell.

Berge, K. L., Evensen, L. S., & Thygesen, R. (n.d.). The Wheel of Writing. A Model of the Writing Domain for the Teaching and Assessing of Writing as a Key Competency. *The Curriculum Journal*.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (3rd ed.). New York: Routledge.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. http://doi.org/10.1177/0265532214542994

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing Assessment in Admission to Higher Education: Review and Framework*. New York: College Entrance Examination Board.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main: Peter Lang.

Holten-Kvistad, A., & Skar, G. (manus). *Utvikling av skriveoppgaver*. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.

Knoch, U. (2007). Do Empirically Developed Rating Scales Function Differently to Conventional Rating Scales for Academic Writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

Kunnskapsdepartementet. (2011). *Motivasjon - Mestring - Muligheter - Ungdomstrinnet*. (St.meld. nr 22 2010-11): Oslo: Det Kongelige Kunnskapsdepartementet.

Linacre, J. M. (2013). *A user's guide to FACETS. Rasch-model computer programs. Program manual 3.71.0*. Hämtad 2015-04-07. Retrieved from http://www.winsteps.com/a/Facets-ManualPDF.zip

Linacre, J. M. (2014). Facets® (version 3.71.4) [Computer Software]. Beaverton, Oregon:

Winsteps.com.

Matre, S., Berge, K. L., Evensen, L. S., Fasting, R. B., Solheim, R., & Thygesen, R. (2011). *Developing National Standards for the Teaching and Assessment of Writing. Rapport frå forprosjekt Utdanning 2020.* Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement*, *46*(4), 371–389. Retrieved from http://www.jstor.org/stable/25651523

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Schumacker, R. E., & Smith, E. V. (2007). Reliability. A Rasch Perspective. *Educational and Psychological Measurement*, *67*(3), 394–409. http://doi.org/10.1177/0013164406294776

Skar, G., Evensen, L. S., & Iversen, J. M. (2015). *Læringsstøttende prøver i skriving 2014. Teknisk rapport.* Trondheim: Nasjonalt senter for skriveopplæring og skriveforskning.

Skar, G., & Iversen, J. M. (2015). *Læringsstøttende prøver i skriving 2014. Teknisk rapport avseende pilotuppgifter HT 2014*. Trondheim: Nasjonalt senter for skriveopplæring og skriveforskning.

Woods, A., & Baker, R. (1985). Item response theory. *Language Testing*, *2*, 117–140. http://doi.org/10.1177/026553228500200202

Appendix

Tolkning av d-värden

Ett enkelt sätt för att förstå vad en given effektstorlek innebär är att konsultera nedanstående uppställning. I denna tabell återges andel av jämförelsegrupp som får lägre poäng än genomsnittseleven i gruppen med högst medelvärde. Ett exempel: om skillnaden mellan flickor och pojkar motsvarar d = 0,90 innebär det att 82 % av pojkarna får lägre resultat än den genomsnittliga flickan.

Effektstorlek	Andel av jämförelsegrupp som får lägre poäng än genomsnittseleven i gruppen med högst medelvärde.
0,0	50 %
0,1	54 %
0,2	58 %
0,3	62 %
0,4	66 %
0,5	69 %
0,6	73 %
0,7	76 %
0,8	79 %
0,9	82 %
1,0	84 %
1,2	88 %
1,4	92 %
1,6	95 %
1,8	96 %
2,0	98 %
2,5	99 %
3,0	99,9 %

Källa: Coe, R. (2002). It's the Effect Size, Stupid. What effect size is and why it is important. Paper presented at the British Educational Research Association annual conference, Exeter, 12-14 September, 2002. Hämtad 2015-01-12 på: http://www.cem.org/attachments/ebe/ESguide.pdf