NTNU

Skrivesenteret
Nasjonalt senter for skriveopplæring og skriveforsking

Teknisk rapport: Pilotering av uppgifter till den nasjonale utvalgsprøven i skriving 2016

G. B. Skar
J. M. Iversen

Innehåll

För att citera denna rapport:
Skar, G. B. & Iversen, J. M. (2016). *Teknisk rapport: Pilotering av uppgifter til den nasjonale utvalgsprøven i skriving.* Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.

Sammanfattning

- Denna rapport beskriver arbetet med pilotering av sex uppgifter till *Den nasjonale utvalgsprøven skriving.*

- Utgångspunkten för rapporten har varit frågan: är det rimligt att anta att uppgifterna kan fungera i utvalgsprøven?

- Resultaten visar att fyra av sex uppgifter överlag får godkänt betyg av lärarna. Det gäller P1553 (5. trinn, skrivhandling: beskriva), P1556b (5. trinn, skrivhandling: överbevisa) P1583 (8. trinn, skrivhandling: beskriva) och P1586 (8. trinn, skrivhandling: överbevisa). De resterande två uppgifterna, P1556b (5. trinn, skrivhandling: överbevisa) och P1584 (8. trinn, skrivhandling: utforska), möter däremot kritik.

- Bedömarpanelen är generellt positiva till uppgifterna. Ingen uppgift på 5. trinn tycks vara särskilt problematisk, men två uppgifter på 8. trinn uppfattas delvis som bristfälliga. Det gäller P1584 och P1583.

- Analysen av genomförda bedömningar visar att det finns indikationer på en en-faktormodell och att data passar Rasch-modellen väl. Analysen visar också att samstämmigheten mellan bedömare för första gången i Skriveprøvens historia med god marginal ligger över den kritiska gränsen på ,70.

- Själva resultaten pekar på att eleverna i genomsnitt presterar strax under skalans mittnivå.

- Undersökning kan inte ge ett entydigt svar på frågeställning. Den huvudsakliga orsaken är svaret varierar med uppgift. Vår rekommendation är därför att vissa uppgifter blir kvar i systemet, medan de andra utgår.

- Vi föreslår att följande blir kvar: P1553, P1556a och P1586. Vi föreslår att följande uppgifter utgår till följd av kritik från lärare eller panel eller avvikande prestationsmönster: P1556b, P1583 och P1584.

1 Inledning och frågeställning

Denna rapport beskriver arbetet med pilotering (Pilot 15) av sex uppgifter till *Den nasjonale utvalgsprøven skriving* (Utvalgsprøven). Utvalgsprøven genomförs varje höst på ett statistiskt representativt urval elever från 5. och 8. trinn. Piloteringen bygger på ett likaledes representativt urval. Den genomförs vid samma tidpunkt och avser nästkommande års utvalsgprøve. På grund av en förändrad design på Utvalgsprøven, som träder i kraft hösten 2016, är detta den sista pilot-rapporten.

För att utvalgsprøven skall fungera optimalt prövas uppgifter ut på förhand. Denna utprövning renderar kvalitativ och kvantitativ data, som analyseras för att svara på frågan: är det rimligt att anta att uppgifterna kan fungera i utvalgsprøven? Om svaret är ja kan uppgifterna fortsatt vara en del av systemet och bli aktuella för senare utvalsgprøver.

Rapporten är uppbyggd på följande sätt. Inledningsvis presenteras en koncis historik och översikt över framtiden för Utvalgsprøven och den besläktade Skriveprøven. Därefter presenteras de sex uppgifter som piloterades. Avsnitten därefter beskriver material och metod för analys av uppgifterna, vilket följs av resultatsammanställning. Den avslutande delen av rapporten sammanfattar resultaten och kommer med några rekommendationer.

1.1 Skriveprøver i Norge – en koncis historik och planer för framtiden

I samband med införande av den senaste läroplanen, *LK06*, beslutades att några av de grundläggande färdigheterna (skrivning, läsning och räkning) skulle bli föremål för nationella populationsprov. Redan efter första försöket skrinlades dock projektet med nationella prov i skrivande, eftersom granskningar visade att reliabiliteten var alltför låg.

År 2010 fick det nationella centret för skrivande (Skrivesenteret) i uppdrag av Kunnskapsdepartementet och Utdanningsdirektoratet att utveckla ett urvalsprov i skrivande (utvalgsprøven i skriving). Utvalgsprøven skulle genomföras av ett nationellt representativt urval elever i början av årskurs 5 och årskurs 8. Elever i respektive årskurs skulle genomföra olika uppgifter. Elevtexterna skulle bedömas av expertbedömare, med särskilda skalor för de olika årskurserna.

Vidare skulle Utvalgsprøven ligga till grund för en så kallad læringsstøttende prøve i skriving – Skriveprøven. Skriveprøven skulle fungera som en pedagogisk resurs med uppgifter som lärare kunde använda för att undersöka skrivförmågan i den egna klassen. Klassens resultat kunde sedan jämföras med resultaten på utvalgsprøven, som genomförts ett år tidigare.

År 2014 lanserades de första skriveprøvene i Utdanningsdirektoratets prøvebank. Proven innehöll uppgifter som genomförts som utvalsgprøve. Detsamma skedde året därpå och är också planerat att ske tidig höst år 2016.

Under år 2015 initierade Skrivesenteret ett utvecklingspaket, som innebär förfinade statistiska metoder, att eleverna skriver två uppgifter (i stället för en) och att alla elever – oberoende av årskurstillhörighet – löser samma uppgifter och bedöms med samma skala. Dessutom innebä r paketet användning av så kallade ankaruppgifter, för att möjliggöra ändamålsenliga skaljusteringar och analyser av uppgifters faktiska svårighetsgrad.

Utvecklingspaketet innebär att utvalgs- och skriveprøven förändras stegvis. Under 2016 kommer varje elev att lösa 2 av totalt 7 tillgängliga uppgifter. De 7 uppgifterna kommer att distribueras på ett sådant sätt att elever från båda årskurser och bedömare blir länkade till varandra. Av de 7 uppgifterna kommer 4 att lanseras som skriveprøve år 2017, medan 3 av dem blir föremål för nya utvalsgprøver i åren som kommer (om detta: se kommande teknisk rapport). Det nya förfarandet innebär också att pilotproven, som är föremål för denna rapport, utgår till förmån för en utökad ”pre-pilot”.

2 Material & metod

Sex uppgifter ingick i 2015 års pilotprov. Totalt deltog 604 elever från 53 skolor, som formade ett nationellt representativt urval, som Skrivesenteret tillhandahållit från Utdanningsdirektoratet. Fördelningen av skolor och antal elever framgår av tabell 2.1.

Tabell 2.1. Uppgifterna fördelade på antal skolor och antal elever

	5. trinn				**8. trinn**			
	P1553	P1556a	P1556b	Totalt	P1583	P1584	P1586	Totalt
Skolor (*n*)	9	10	8	27	8	10	8	26
Elever (*n*)	101	101	99	301	101	101	101	303

Förutsättningarna för deltagande i Pilot 2015 var desamma som vid tidigare år. Provet genomfördes under veckorna 35 och 36. Varje elev löste en uppgift och varje text blev bedömd av två oberoende bedömare. Varje elev fick 45 minuters skrivtid. Elever i 5. trinn skrev för hand, medan elever i 8. trinn skrev på dator. I tillägg till uppgiftsformuleringarna försågs lärare med detaljerade instruktioner om tillvägagångssätt för uppgiftslösning och med information om den konstruktförståelse som ligger bakom uppgifterna. Instruktionerna innehöll också s.k. støttespørsmål, som var ägnade åt användning i ”idémyldringsøkten”, som enligt instruktion skall genomföras innan eleverna skriver.[1] I figur 2.1 nedan återges de sex uppgifterna.

[1] Fullständiga lärarhandledningar tillhandahålls av Skrivesenteret efter beslut från Utdanningsdirektoratet.

5. trinn	Uppgift P1553	Uppgift P1556a	Uppgift P1556b
	Hva er spesielt med hjemstedet ditt? En familie med barn på din alder skal flytte dit. De har aldri vært der før, og de vil gjerne lese om stedet på forhånd. Skriv en tekst der du beskriver hjemstedet ditt for en familie som skal flytte dit.	En hjelm kan bidra til å minske skade ved sykkelulykker. Likevel bruker bare litt over halvparten av de voksne hjelm når de sykler. Skriv en tekst der du overbeviser voksne syklister om at de må bruke sykkelhjelm. Teksten din skal trykkes i et hefte om sikkerhet i trafikken som skal deles ut til alle foreldre på skolen.	Foreldrene på skolen har foreslått at ingen av barna skal få lov til å feire Halloween i år. Noen av grunnene til dette er at det kan bli bråk, at ungene spiser for mye godteri og at noen følte seg utestengt i fjor og ble lei seg. Du vil prøve å påvirke foreldregruppa til å forandre mening. Dette vil du gjøre ved å skrive til dem for å overbevise dem om at dere skal få lov til å feire Halloween. Skriv en tekst der du overbeviser foreldregruppa om at dere skal få lov til å feire Halloween.
8. trinn	Uppgift P1583	Uppgift P1584	Uppgift P1586
	Digitale verktøy er en del av hverdagen vår. Ulike aldersgrupper bruker datateknologi på forskjellige måter. Besteforeldrene dine er interesserte i å vite mer om hva du bruker digitale verktøy til i skolearbeid og fritid. Skriv en tekst der du beskriver hva du bruker digitale verktøy til i skolearbeid og fritid. Tenk deg at bestemor og bestefar skal lese teksten.	For 65 millioner år siden skjedde noe som utryddet over halvparten av alle arter som levde på jorda. Dinosaurene, som hadde levd på jorda i 160 millioner år, forsvant gradvis. Dette er en av naturhistoriens største gåter: hvorfor døde dinosaurene ut? Skriv en tekst der du utforsker hvorfor dinosaurene døde ut. Du skriver teksten til et hefte klassen din lager om utdødde plante- og dyrearter.	En barneskole i nærheten skal gjennomføre en kampanje for å få elevene på femte trinn til å bruke sykkelhjelm. Skriv en tekst der du overbeviser femteklassingene om at de bør bruke sykkelhjelm.

Figur 2.1. Uppgifter som är piloterade

Som framgår av uppställningen fördelar sig uppgifterna på olika skrivhandlingar (jfr Berge, Evensen, & Thygesen, 2016; Skar, Evensen, & Iversen, 2015), något som är i

linje uppdraget från Utdanningsdirektoratet. Värt att notera är att uppgift P1584 skiljer sig en aning från övriga uppgifter. På grund av tematiken beslutades att i lärarinstruktionerna inkludera en "ingångstext" (se bilaga 1), som skulle fungera som gemensamt kunskapsunderlag. Ingångstexten lästes högt av läraren före provtillfället och fanns i övrigt inte tillgänglig för elever.

För att undersöka hur väl uppgifterna kan antas fungera i utvalgsprøven samlades olika typer av material in. För det första deltog alla lärare, som administrerat proven, i en enkätundersökning, som bland annat fokuserade på lärarens upplevelse av tematikens relevans, instruktionernas klarhet och elevernas motivation (se resultatkapitel). Också medlemmarna av bedömarpanelen deltog i en enkätundersökning som fokuserade på uppfattningar av uppgifternas kvalitet (se resultatkapitel). För det andra genomförde Skrivesenterets medarbetare 10 observationer av genomförandet av Pilot 15 (för observationsschema, se bilaga 2). För det tredje genomfördes i februari 2016 bedömningar av de insamlade elevsvaren. Under en samling för bedömarpanelen bedömdes varje elevtext i sex kategorier (kommunikation, innehåll, textstruktur, språkbruk, stavning och interpunktion) av två av varandra oberoende bedömare. I tabell 2.1 framgår det insamlade materialet.

Tabell 2.1. Insamlat material.

	Lärarenkäter	Panelenkäter	Observationer	Elevtexter	Bedömningar
5 trinn	24	31	4	301	4 266
8 trinn	23	27	6	303	4 163

Av redovisningen ovan framgår att Skrivesenteret har tillgång till ett brett och varierat material, vilket ökar chansen att kunna göra plausibla antaganden om uppgifternas kvalitet.

2.1 Analysmetod

I huvudsak har materialinsamlingen inneburit tre typer av material. Det första består av svar på två enkätundersökningar, riktade till lärare respektive expertbedömare. Svaren har analyserats med hjälp av deskriptiv statistik och i resultatavsnittet återges svarsfrekvenser och andelar för olika svarsalternativ. Frågor och svarsalternativ återges i resultatkapitlet av denna rapport.

Det andra materialet består av fältnotat från observationer. Dessa återges i förkortad form i resultatkapitlet längre fram.

Det tredje materialet består av elevtextbedömningar. Liksom vid två tidigare tillfällen (Skar & Iversen, 2015, 2016) har bedömningsunderlaget analyserats på tre sätt. För det första har faktoranalyser (PCA-analyser) genomförts i syfte att undersöka underliggande strukturer. För det andra har bedömarreliabiliteten skattats med hjälp av tekniker från *classical test theory*. I rapporten redovisas, som i tidigare fall, *intra class correlation coefficient*. För det tredje har vi använt oss av *many-facet Rasch-measurement* (MFRM) för att identifiera oregelbundenheter i bedömningen och för att ta fram

elevresultat i vilka vi kontrollerat för bedömares stränghet. I tabell 2.1 framgår att 8 429 bedömningar ligger till grund för analysen. Bedömningarna har genomförts har på 5. och 8. trinn genomförts av 33 respektive 32 bedömare. Varje bedömare har således genomfört i snitt 130 bedömningar.

2.1.1 MFRM-analys

MFRM-analyser genomförs i datorprogrammet *Facets* (Linacre, 2014), som inkluderar en utbyggd version av den så kallade Rasch-modellen (Rasch, 1980). Den ursprungliga, dikotoma Rasch-modellen gör gällande att sannolikheten för ett korrekt eller felaktigt svar är en funktion av differensen mellan en testtagares förmåga och ett items svårighet. Enkelt uttryck ökar sannolikheten att en elev svarar rätt ju större avståndet är mellan elevens förmåga och uppgiftens svårighetsgrad (för ingående beskrivningar, se t.ex. Bond & Fox, 2015; Engelhard, 2013). Användningen av Rasch-modellen i dess ursprungliga eller utvidgade form bygger på en acceptans av detta som en rimlig premiss.

En förutsättning för meningsfulla mätningar är att skattningen av en elevs förmåga är item-invariant och vice versa, att skattningen av ett items svårighet är person-invariant (Engelhard, 2013). I otekniska termer betyder det att estimatet av en persons förmåga gäller oaktat vilket item denne stöter på, liksom estimatet av ett items svårighet gäller oaktat vilken person som tar det. När så är fallet kommer det alltid vara mer sannolikhet att en person med hög förmåga klarar ett givet item än att en person med låg förmåga gör det. Översatt till skrivbedömning betyder det att en elev med god skrivförmåga sannolikt får högre poäng på en given uppgift, ett givet område av en given bedömare än en elev med sämre skrivförmåga. När så inte är fallet, när till exempel personer med liten förmåga har större sannolikhet att klara ett svårt item, än personer med hög förmåga, är risken stor att detta item mäter något annat. Rasch-modellen används för att ta fram skattningar av personförmåga och itemsvårighet som är invarianta.

Rasch-analyser har också andra fördelar. Bland annat kan nämnas att analysen innebär att råpoäng transformeras till så kallade logit-värden. När data-setet passar Rasch-modellen innebär transformeringen att skalan går från ordinal till intervall. En intervallskala innebär i sin tur att avstånden är ekvidistanta; skillnaden mellan 1 och 2 är lika stor som den mellan 3 och 4. En annan, och för Skriveprøven önskvärd effekt, är att Rasch-analysen bibringar oberoende estimat av t.ex. personer och bedömare. Analysen kan alltså på frågan: vilken skrivförmåga har den här eleven, oberoende av vem i bedömarpanelen som bedömer honom/henne?

Den ursprungliga Rasch-modellen är avpassad för dikotoma items, men i fallet med skrivbedömning spelar dock andra faktorer roll, så som exempelvis bedömares stränghet. Därför finns behov för en modell som kan ta hänsyn också till sådana aspekter (Barkaoui, 2014; Eckes, 2015). Linacre (1989, citerad i Linacre, 2013) har utvidgat modellen för att kunna ta hänsyn till flera aspekter som kan antas vara av relevans, t.ex. bedömare, tidpunkt för provgenomförande och liknande. I Pilot 15 utgår

vi från att sannolikheten (*P*) att person *n* (t.ex. Janne), på uppgift *m* (t.ex. P1584), på bedömningsområde *i* (t.ex. rettskriving) av bedömare *j* (t.ex. Ole) tilldelas *k* (t.ex. 3) relativt att samma person under samma förutsättningar tilldelas *k*-1 (t.ex. 2) är en funktion av flera faktorer. Den modell som använts i Pilot 15 ser ut såhär:

log(Pnmijk/Pnmij(k-1)) = Bn - Am - Di - Cj - Fk,

där Bn avser elevens skrivförmåga, Am är uppgiftens svårighet, Di är bedömningsområdets svårighet, Cj är bedömarens stränghet och Fk svårigheten att placeras i kategori k relativt till att placeras i kategori k-1.

Aspekterna som tas hänsyn till i MFRM-modellen kallas *facets*. I Pilot 15 opererar vi med fyra facets: personer (elever), uppgifter, bedömningsområden och bedömare. Enskilda elever, uppgifter, områden och bedömare kallas *element*. Varje element erhåller genom analysen ett estimat på förmåga eller svårighet.

I huvudanalysen genereras estimat som bygger på all tillgänglig data. För elever blir ett sådant estimat detsamma som ett summerat resultat (ett sumskåre).[2] Liksom tidigare har vi också tagit fram två typer av estimat, som representerar funktionskompetenserna (dvs. kommunikation, innehåll, textstruktur och språkbruk) och kodningskompetenserna (dvs. stavning och interpunktion). Dessa estimat har tagits fram i analyser där bedömares logit-värden och Rasch-Andrich Thresholds från huvudanalysen varit ankrade (om ankring, se t.ex. Bond & Fox, 2015).

En viktig förutsättning för att *Facets* skall kunna generera meningsfulla estimat är att alla element är länkade till varandra. I Pilot 15 gjordes det på samma sätt som i Utvalgsprøven 15, nämligen genom länktexter (Skar & Iversen, 2016). Dessutom bedömdes varje elevsvar av två av varandra oberoende bedömare. Texter från de olika skolorna distribuerades jämt över de olika paren.

En annan viktig förutsättning för meningsfulla estimat är den som relaterar till *model fit*. När de modellförväntade resultaten skiljer sig mycket från de observerade passar inte data modellen. Det kan finnas olika orsaker till det, t.ex. att de items (bedömningsområden) som ingår inte bidrar till mätning av *ett* underliggande konstrukt, eller att bedömarna varit alltför oense.

För att undersöka kravet om att enskilda items skall bidra till mätning av ett konstrukt kan man genomföra faktoranalyser, vilket vi gjort (se kapitel 3.3). Man kan också undersöka enskilda items med hjälp av så kallad fit-statistik. Skattningar av det senare rapporteras i *Facets* i infit- och outfit-värden för element, facets och för data-setet som helhet (och kallas då "global fit"). För bedömning av hela data-setet är det emellertid mer rimligt att undersöka absoluta standardiserade residualer (Eckes, 2015). En vedertagen riktlinje säger att data passar modellen bra om bara en

[2] Övriga element får också estimat. Det gäller dock ej uppgifter, eftersom elever bara skrivit en uppgift, vilket medför att data-seten inte är överlappande.

viss andel 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 % av observationerna har en standardiserad residual över 3,0.

Om data bedöms passa modellen finns en rad andra, nyttiga värden att ta hänsyn till. Dessa anger hur mätningen lyckats separera element i olika nivåer av färdighet (för personer), svårighet (för items) och stränghet (för bedömare), samt med vilken precision. De anger för bedömar-facet också andel fullständig samstämmighet och korrelation med övriga bedömare. För skalorna, slutligen, anges huruvida skalstegen följer ett förväntat mönster, nämligen att högre förmåga innebär större sannolikhet att observeras i högre kategorier.

I figur 2.2 återges de kvalitetsindikatorer som är intressanta att studera, och som redovisas i resultatkapitlet, i en version som är bearbetad utifrån Skar och Iversen (2016) .

	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 (outliers). 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.
	ASR	Standardiserade residualer	Globalt estimat av huruvida data passar modellen. Riktlinjer: max 5 % av observationerna har en standardiserad residual > 2,0; max 1 % av observationerna standardiserad residual > 3,0.
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-modellens förväntningar.
Skalor	Rasch-Andrich Thresholds		Värdet anger var på logitskalan som två kategorier är lika sannolika; skall öka med kategorier
	Deskriptiv statistik		Anger fördelning på de olika skalstegen i antal och andel. Varje kategori/skalsteg bör vara observerat minst 10 tillfällen.

Figur 2.2. Kvalitetsindikatorer från MFRM-analyser.

3 Resultat

I det här kapitlet presenteras resultaten från de tre delundersökningarna. Först återges resultatet av de två enkätundersökningar som genomförts. I den första har lärare lämnat synpunkter på olika aspekter av uppgifterna. I den andra har medlemmar ur bedömningspanelen angett huruvida uppgiften fungerar eller inte. Därefter redovisas resultaten av de observationer som Skrivesenteret genomförde. Slutligen presenteras analysen av bedömningsdata.

3.1 Lärares och panelens återkoppling

De lärare som administrerade Pilot 15 ombads bedöma 10 påståenden om bland annat uppgifterna, lärarinstruktionerna och elevernas motivation. Resultaten för de sju första påståendena redovisas uppdelat på uppgift, medan en sådan uppställning inte är gjord i samband med de tre sista påståenden som är mer allmänna till sin karaktär.

Det första påståendet löd: "Skriveoppgaven gir eleven mulighet til å vise skriveferdighetene sine". Resultatet återfinns i tabell 3.1, som visar att en övervägande del av informanterna (38 av 47) menade att påståendet stämmer, eller stämmer mycket väl. Två lärare menade att påståendet inte stämmer, vilket är en invändning som kan vara viktig. Dessa två lärare hade genomfört P1584, dvs. uppgiften om dinosaurer, som inkluderade en "ingångstext". Ytterligare sju lärare (varav en som genomfört P1584), menade att påståendet stämde någorlunda väl. Uppgiften P1556b dominerade här (3 svar).

Tabell 3.1. Skriveoppgaven gir eleven mulighet til å vise skriveferdighetene sine

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	0	0	0	2	0	2
Stemmer nokså godt	1	1	3	1	1	0	7
Stemmer godt	4	4	4	4	3	4	23
Stemmer svært godt	2	5	0	2	3	3	15
Tot.	7	10	7	7	9	7	

Det andra påståendet löd: "Den muntlige instruksjonen som læreren skal lese opp til skriveoppgaven er forståelig". Också här menade lärarna till övervägande del att påståendet stämde gott, eller mycket gott. Ingen lärare menade att påståendet inte stämde alls, men sex lärare – fördelade på uppgifterna P1556a, P1583 och P1584, menade att påståendet stämde bara någorlunda väl. Detaljerna redovisas i tabell 3.2.

Tabell 3.2. Den muntlige instruksjonen som læreren skal lese opp til skriveoppgaven er forståelig

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	0	0	0	0	0	0
Stemmer nokså godt	0	2	0	1	3	0	6
Stemmer godt	4	2	7	2	3	1	19
Stemmer svært godt	3	6	0	4	3	6	22
Tot.	7	10	7	7	9	7	

Det tredje påståendet löd: "Den skriftlige instruksjonen som eleven får til skriveoppgaven er forståelig". Av 47 informanter var 37 eniga i större eller mindre utsträckning, medan 10 reserverade sig. En lärare som genomfört P1556a menade att påståendet inte stämde. Bland de lärare som menade att påståendet stämde någorlunda väl finns uppgifterna P1556a, P1556b, P1583 och P1584 representerade. Resultaten presenteras i tabell 3.3.

Tabell 3.3. Den skriftlige instruksjonen som eleven får til skriveoppgaven er forståelig

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	1	0	0	0	0	1
Stemmer nokså godt	0	1	2	3	3	0	9
Stemmer godt	6	4	5	0	6	4	25
Stemmer svært godt	1	4	0	4	0	3	12
Tot.	7	10	7	7	9	7	

Det fjärde påståendet löd: "Den skriftlige instruksjonen som eleven får til skriveoppgaven er forståelig". Tämligen många lärare (16 av 47) menade att detta stämde bara någorlunda väl. Dessa lärare representerade samtliga uppgifter, även om svar som representerar P1556b och P1584 dominerar. Två lärare (P1556a och P1584) menade sig vara oförmögna att alls instämma i påståendet. Resultaten redovisas i tabell 3.4.

Tabell 3.4. Elevene kommer raskt i gang med å løse oppgaven

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	1	0	0	1	0	2
Stemmer nokså godt	2	1	6	1	5	1	16
Stemmer godt	5	4	1	5	1	5	21
Stemmer svært godt	0	4	0	1	2	1	8
Tot.	7	10	7	7	9	7	

Den femte påståendet löd: "Temaet for oppgaven er relevant". Av 11 lärare som inte instämt i påståendet representerar nio av dem uppgifterna P1556b och P1584. Också uppgifterna P1556a och P1586 finns representerade. Övriga informanter menade att påståendet stämde gott eller mycket gott. Resultaten redovisas i tabell 3.5.

Tabell 3.5. Temaet for oppgaven er relevant

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	0	1	0	1	0	2
Stemmer nokså godt	0	1	3	0	4	1	9
Stemmer godt	4	2	3	3	3	4	19
Stemmer svært godt	3	7	0	4	1	2	17
Tot.	7	10	7	7	9	7	

Det sjätte påståendet löd: "Elevene mine var motiverte for å skrive en tekst om dette temaet". En majoritet av informanterna menade att påståendet stämde gott eller mycket gott, men 20 informanter hade svårigheter med att instämma delvis eller alls. En majoritet av dessa, i sin tur, representerade uppgifterna P1556b och P1584. Detaljerna redovisas i tabell 3.6.

Tabell 3.6. Elevene mine var motiverte for å skrive en tekst om dette temaet

	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	0	0	0	4	0	4
Stemmer nokså godt	2	3	5	2	3	1	16
Stemmer godt	5	3	2	3	2	6	21
Stemmer svært godt	0	4	0	2	0	0	6
Tot.	7	10	7	7	9	7	

Det sjunde och sista uppgiftsspecifika påståendet löd: "Det er satt av nok tid til å løse oppgaven". Tolv lärare hade svårt att instämma i detta påstående. Återigen finns P1584 representerad, liksom P1556b, men det gör övriga uppgifter också, om än med 1 eller 2 informanter. Resultaten presenteras i tabell 3.7.

Tabell 3.7. Det er satt av nok tid til å løse oppgaven

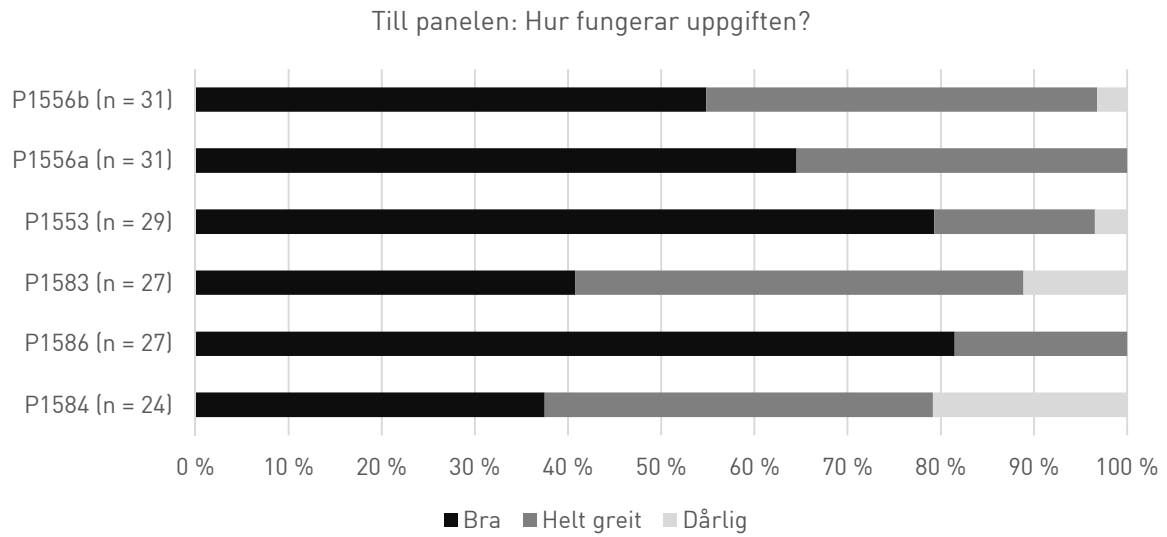
	P1553	P1556a	P1556b	P1583	P1584	P1586	Tot.
Stemmer ikkje	0	1	0	0	2	0	3
Stemmer nokså godt	1	0	2	2	3	1	9
Stemmer godt	4	3	4	3	0	3	17
Stemmer svært godt	2	6	1	2	4	3	18
Tot.	7	10	7	7	9	7	

Tabell 3.8 presenterar resultaten relaterade till följande tre påståenden: (1) "Lærerinstruksen er passe lang", (2) "Lærerinstruksen forklarer godt hvordan jeg skal gjennomføre prøven" och (3) "Materiellet som helhet er oversiktlig". I tabellen framkommer att en övervägande majoritet av informanterna menar att detta stämmer gott eller mycket gott. Detta indikerar att materialet som sådant i de allra flesta fall vållar få bekymmer.

Tabell 3.8. Tre uppgiftsoberoende påståenden

	Stemmer ikkje	Stemmer nokså godt	Stemmer godt	Stemmer svært godt	Tot.
Påstående 1	1	7	22	17	47
Påstående 2	1	2	24	20	47
Påstående 3	2	2	27	16	47

Medlemmar ur bedömarpanelen fick, i samband med att panelen träffades för seminarium, frågan: "Hur fungerar uppgiften?" Panelmedlemmarna instruerades att ta hänsyn till om eleverna genomförde den uppgift som uppgiftformuleringen angav. En uppgift som fungerar dåligt blir därmed en där elever skriver annat än det som uppgiften efterfrågat. Resultatet av undersökningen presenteras i figur 3.1



Figur 3.1. Bedömarpanelens skattningar av uppgifternas kvalitet.

Som framgår av figuren anser en stor andel av de tillfrågade panelmedlemmarna att uppgifterna fungerar bra eller okej. Två uppgifter skiljer ut sig, nämligen P1583 och P1584 som drygt 10 % respektive drygt 20 % av panelmedlemmarna inte tycker fungerar.

3.2 Skrivesenterets observationer av genomförande

Skrivesenterets personal har genomfört 10 observationer av genomförandet av Pilot 15. Dessa observationer har gett tillgång till rik data (jfr observationsschemat), men av utrymmesskäl återges bara de mest centrala resultaten här. Dessa gäller i vilken utsträckning det varit möjligt för läraren att genomföra provet med utgångspunkt i lärarvägledningen, om eleverna tycks ha förstått uppdraget och eventuella kommentarer från läraren om hur uppgiften och/eller genomförandet fungerade. Resultaten återges schematiskt i tabell 3.9.[3]

Tabell 3.9. Sammanfattning av observationer

	Genomförande	Elevers skrivande	Kommentarer
P1553	Läraren, som fram till tio minuter före lektionsstart trodde att Skrivesenteret skulle sköte genomgången, genomförde provet enligt instruktion.	Alla elever skrev; några blev färdiga mycket snabbt, andra skrev tiden ut.	Läraren menade att elevernas motivation varierade och att många elever missförstod uppgiften till att handla om bostaden. Läraren önskade att Skrivesenteret skulle ange genre.
P1556a	*Ej observerad*	*Ej observerad*	*Ej observerad*
P1556b	Skola 1: Läraren genomförde provet enligt instruktion.	Skola 1: Alla elever skrev, men några färdiga redan efter 10 minuter, de flesta efter cirka 35 minuter.	Skola 1: Läraren önskade mer information om vilken hjälp som var tillåten att ge under provet.

[3] P1556a uppgifter har inte observerats. Detta har schematekniska skäl.

	Skola 2: Läraren genomförde provet enligt instruktion. Skola 3: Läraren genomförde provet enligt instruktion.	Skola 2: Alla elever skrev; många elever färdigställde texten ganska snabbt. Skola 3: Eleverna skrev och de flesta var klara efter cirka 35 minuter.	Skola 2: Läraren själv inte begeistrad i temat, men temat engagerade, enligt läraren, eleverna. Skola 3: Läraren menade att uppgiften var avancerad, att eleverna engagerades av provsituationen; läraren önskade mer specifikation om vilka avsteg som var möjliga att göra; äraren önskade en tydligare genrebeställning.
P1583	Skola 1: Läraren genomförde provet enligt instruktion. Skola 2: Läraren genomförde provet enligt instruktion.	Skola 1: Alla elever skrev, men graden av engagemang var liten. Många elever gav uttryck för viss osäkerhet om hur texten skulle skrivas. Skola 2: Alla elever skrev; många elever färdigställde texten ganska snabbt.	Skola 1: Läraren menade att temat mor- och farföräldrar kunde vara känsligt för enskilda elever; läraren önskade en tydligare genrebeställning; läraren önskade att det skulle vara möjligt att använda tavlan under genomgången. Skola 2: Läraren hade önskat att eleverna fick hela instruktionen (och inte bara sista meningen). Läraren trodde att 45 minuter skulle vara för lite tid, men det visade sig vara tillräckligt.
P1584	Läraren missförstod uppmaningen att läsa uppgiften högt en andra gång. Läraren använde inte frågor knutna till "idémyldringen"	Eleverna skrev, men ytterst korta texter.	Läraren menade att skrivehandlingen "utforska" är svår för elever; läraren menade att uppgifterna borde ha "ikke-faglige tema", som alla elever kan något om.
P1586	Skola 1: Läraren genomförde provet enligt instruktion. Skola 2: Läraren genomförde provet enligt instruktion. Skola 3: Läraren genomförde provet enligt instruktion.	Skola 1: Alla elever skrev, några elever skrev hela tiden ut. Skola 2: Eleverna skrev och de flesta var klara efter cirka 35 minuter. Skola 3: Alla elever skrev, de flesta klara före utsatt tid.	Skola 1: Läraren slet med PC Skola 2: Läraren menade att genomförandet av ett dylikt prov kräver minst två lärare närvarande. Skola 3: Läraren tyckte att eleverna var engagerade under genomgången; läraren menade att eleverna kunde få tydligare beställningar (t.ex. vilka rubriker som skall användas)

Genomgången, som i denna form blir tämligen summarisk och något ytlig, visar att lärarinstruktionerna tycks fungera. Lärarna genomför provet på ett sådant sätt som Skrivesenteret avsett. Genomgången visar också att alla observerade elevgrupper lyckas skriva och att tiden verkar vara rimlig. En av de kommentarer som återkommer bland lärarna ger uttryck för önskemål om större grad av specificering hur skrivhandlingarna skall utföras (t.ex. genom att ange genre). Andra kommentarer rör det faktum att läraren inte skall använda tavlan under genomgången och önskemål om ökad tydlighet om på vilket sätt det är möjligt att hjälpa elever under provgenomförandet. Tre lärarkommentarer relaterar till själva uppgiften: en lärare menar att mottagaren i P1583 kan vara känslig för vissa elever, en lärare menar att eleverna behöver möta "ikke-faglige" uppgifter (P1584) och en lärare säger sig själv vara skeptisk till tematiken, men kunde observera att den slog an hos eleverna.

Underlaget från observationerna är av för liten omfattning för att det skall kunna vara möjligt att dra slutsatser om enskilda uppgifters kvaliteter (något som inte heller varit syftet; jfr ovan). Det kan dock vara värt att notera att uppgift P1584 som är den som kritiserats av andra lärare och i högst grad av panelen, också kritiseras här.

Sammantaget visar resultaten från de två enkätundersökningarna och från observationerna följande. De flesta uppgifter bedöms av lärare och panel att fungera gott eller mycket gott. Undantagen gäller P1584, som möter kritik på båda ställen, och P1556b som möter viss kritik från lärarna. Observationerna visar att instruktionsmaterialet fungerar och att elever i alla de tio observerade klasserna oberoende av uppgift kan medverka i provet på ett meningsfullt sätt.

3.3 Bedömningar

Detta delkapitel är tudelat. I det förstone presenteras analyser relaterade till psykometrisk kvalitet. I kapitel 3.3.1 redovisas resultatet av faktoranalysen och i kapitel 3.3.2 resultaten som är relaterade till frågan om data passat MFRM-modellen. I den andra övergripande delen, kapitel 3.3.3, presenteras så själva resultaten.

3.3.1 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 de siste gjennomgangene inkluderer vi først en enkel korrelasjonsmatrise. Vi ser i tabell 3.10 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.

Sammenlignet med høstens gjennomgang ser vi nå omtrent samme mønster. Det er en svak tendens til at de fire første områdene har høyere korrelasjonsverdier i denne gjennomgangen, mens de to siste var sterkere innbyrdes korrelerte i høst. Forskjellen mellom verdiene i de fire første og de to siste er større i denne gjennomgangen.

Tabell 3.10. Korrelasjonsmatrise av vurderingsområdene på 5. trinn

	V1[a]	V2	V3	V4	V5	V6
V1	–					
V2	0,62	–				
V3	0,68	0,65	–			
V4	0,59	0,56	0,68	–		
V5	0,28	0,32	0,39	0,41	–	
V6	0,36	0,35	0,40	0,45	0,49	–

[a]V = vurderingsområde.

Tabell 3.11 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 initiale 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 1, velger vi å gjennomføre en rotasjon med to faktorer

Tabell 3.11. Prinsipal komponentanalyse av vurderingsområdene på 5. trinn

	Faktor 1	Faktor 2
Eigenvalue	3,46 (57 %)	0,95 (16 %)
Kommunikasjon	0,90	-0,09
Innhold	0,86	-0,05
Tekstoppbygging	0,86	0,06
Språkbruk	0,70	0,22
Rettskriving	-0,04	0,90
Tegnsetting	0,07	0,80

Resultatet av analysen er som nevnt at den indikerer en en-faktormodell. Verdiene på eigenvalue og på hvordan de ulike vurderingsområdene lader opp mot den enkelte faktor er tilnærmet lik det som vi så i høst. Også 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 og de tidligere rundene. Faktor 1 forklarer 57 % av variansen, mens faktor 2 forklarer 16 %.

Også på 8. trinn har vi gjennomført samme analyser. Først ser vi fra korrelasjonsmatrisen i tabell 3.12 at mønstret på 5. trinn gjentar seg. Sammenlignet med i høst er korrelasjonsverdiene noe lavere i alle tilfeller. Mønsteret er imidlertid fortsatt det samme. De fire vurderingsområdene kommunikasjon, innhold, tekstoppbygging og språkbruk korrelerer sterkt innbyrdes. Korrelasjonen er også stort sett 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 3.12. Korrelasjonsmatrise av vurderingsområdene på 8. trinn

	V1[a]	V2	V3	V4	V5	V6
V1	–					
V2	0,70	–				
V3	0,70	0,70	–			
V4	0,56	0,63	0,71	–		
V5	0,40	0,44	0,47	0,57	–	
V6	0,43	0,44	0,57	0,62	0,57	–

[a]V = vurderingsområde.

Prinsipal komponentanalysen viser en noe sterkere tendens til en-faktorstruktur enn på 5. trinn. Forskjellen er imidlertid ikke så stor som i høst. Vi har imidlertid også sett denne forskjellen 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 64 % av total varians (69 % høsten 2015). I den andre faktoren med eigenvalue på 0,82 (kun 0,67 høsten 2015) lader de to siste vurderingsområdene sterkt. Vi har likevel inkludert to faktorer i rotasjonen på samme måte som på 5. trinn. Faktorstrukturen er tilnærmet som tidligere og vi ser en relativt klar en-faktorstruktur hvor fire vurderingsområder lader sterkt.

Tabell 3.13. Prinsipal komponentanalyse av vurderingsområdene på 8. trinn

	Faktor 1	Faktor 2
Eigenvalue	3,86 (64%)	0,82 (14 %)
Kommunikasjon	0,95	-0,09
Innhold	0,90	-0,02
Tekstoppbygging	0,77	0,20
Språkbruk	0,46	0,52
Rettskriving	-0,06	0,91
Tegnsetting	0,04	0,85

Prinsipal komponentanalysen gir også i denne runden dekning for en en-faktorstruktur. Som tidligere er det en annen faktor som i stor grad lades opp av tegnsetting og

rettskriving, men denne faller i utgangspunktet under en cutoff på 1. Innhold, kommunikasjon, språkbruk og tekstoppbygging ser ut til å lade sterkt opp mot faktor 1.

3.3.2 MFRM-analys

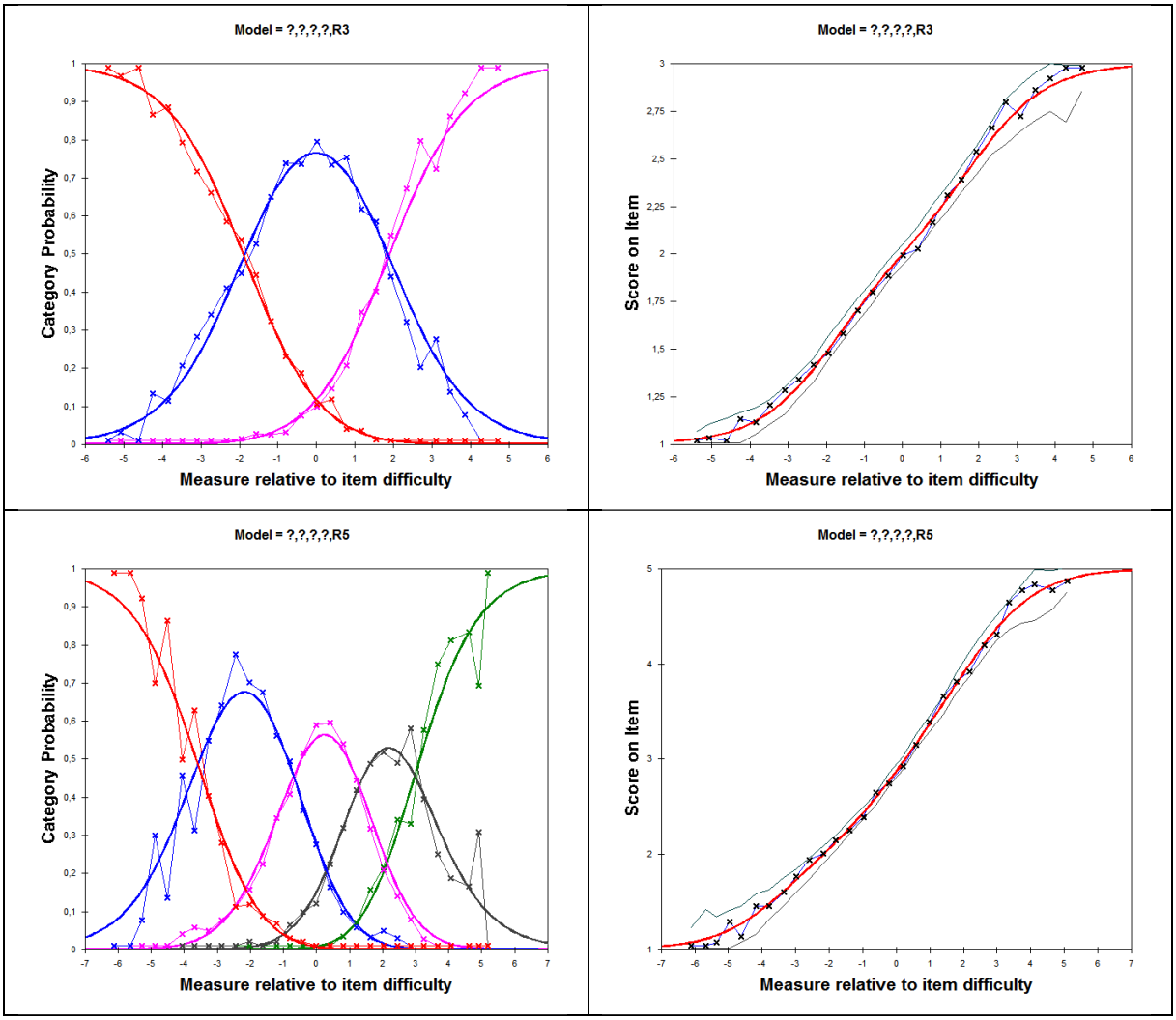
I tabell 3.14 redovisas antalet och andelen standardiserade residualer. Analysen visar att data passar Rasch-modellen väl, något som stämmer överens med erfarenheter från tidigare provomgångar (Pilot 14, Utvalgsprøven 15). Andelen residualer håller sig inom marginalen, oaktat de är 2,0 eller över eller 3,0 eller över.

Tabell 3.14. Standardiserade residualer

	Res ≥ 2,0[a]		Res ≥ 3,0	
	Antal	Andel (%)	Antal	Andel (%)
5. trinn	202	4,7	9	0,2
8. trinn	186	4,8	22	0,5

[a]Res = standardiserade residualer.

I figur 3.2 redovisas en panel av grafer som också är relaterade till *global fit*. Bilderna kan tolkas såhär. Ju bättre överensstämmelse det är mellan de heltecknade färgade linjerna (modellen) och linjerna med kryss (empiriska observationer), desto bättre passar data modellen. Vi kan se att graferna bekräftar bilden vi från från undersökningen av standardiserade residualer. Vi ser också att i områden med få observationer (i ytterkanten på skalorna) är skillnaden mellan observationer och modell något större än i övrigt.



Figur 3.2. Panelen består av fyra figurer, som illustrerar hur väl data passar modellen. Panelens två översta grafer illustrerar analysen av data från 5. trinn, medan de två understa illustrerar analysen av data från 8. trinn. Graferna visar att resultatet i stort sett är tillfredsställande, även om konfidensintervall är bredare och avsteget från modellen större i skalornas ytterkant.

I tabell 3.15 presenteras mått på de kvalitetsindikatorer som presenterades i figur 2.2 (infit och outfit undantaget), nämligen homogenitetsindex (Q), strata (H) och reliabilitet (R) samt för bedömarna procent enighet och den genomsnittliga korrelationen (RMSE och True S.D. används vid beräkningen av H och R).

Tabell 3.15. Kvalitetsindikatorer

		RMSE	True S.D.	Q (df)	H	R	Enig %	SR-ROR
5 tr.	Elever	,69	1,91	2794** (300)	3,99	,88	–	–
	Skalor	,08	0,29	74,4** (5)	5,29	,92	–	–
	Bedömare	,18	0,52	292,3** (32)	4,14	,89	62,5	,39
8 tr.	Elever	,47	1,86	4330,8** (302)	5,58	,94	–	–
	Skalor	,06	0,33	162,6** (5)	7,23	,96	–	–
	Bedömare	,13	0,44	367,9** (31)	4,70	,91	42,6	,43

Notera. Skalor = Skalor/bedömningsområden; RMSE = Root Mean Square Error; True S.D. = sann varians; Q = homogenitetsindex; df = frihetsgrader; H = separationsindex; R = reliabilitetsindex; Enig % = procent fullständig samstämmighet; SR-ROR = Single Rater-Rest of Raters. ** p = < ,01.

Resultaten för 5. trinn indikerar att det finns upp till fyra statistiskt distinkta klasser av skrivförmåga. Reliabiliteten är emellertid inte helt övertygande (*R* = ,88) och lägre än för Utvalgsprøven 15 (*R* = ,94). På 5. trinn används emellertid en skala med tre steg, och talen indikerar att det är fullt möjligt att separera eleverna i tre nivåer. Resultaten för 8. trinn visar att det är möjligt att fästa större tilltro till inledningen i förmågenivåer (*H* = 5,58, *R* = ,94). Detta kan jämföras med *H* = 6,94 och *R* = ,96 för Utvalgsprøven 15.

Analysen av skalorna (bedömningsområdena) på 5. trinn visar att dessa skiljer sig åt i svårighet, uppemot fem nivåer. I analysen av Utvalgsprøven 15 registrerades 6 klasser med hög reliabilitet (,96). Skalanalysen på 8. trinn visar också skillnad mellan bedömningsområdena. Resultaten indikerar att det finns cirka 7 svårighetsnivåer kopplade till skalorna och reliabiliteten är hög (*R* = ,96). Ännu fler nivåer registrerades dock i Utvalgsprøven 15 (*H* = 9,57; *R* = ,98).

Bedömarna på 5. trinn uppvisar färre klasser av bedömarstränghet än vad som var fallet i Utvalgsprøven 15. För Pilot 15 gäller att vi någorlunda säkert kan separera bedömarna i fyra klasser av stränghet (*H* = 4,14; *R* = ,89), vilket är en markant skillnad mot Utvalgsprøven 15 (*H* = 6,18; *R* = ,95). Resultaten visar att bedömarna alltså bedömer mer likt varandra, vilket också märks när vi studerar enighetsprocenten (62,5 % mot 57,9 % i Utvalgsprøven 15) och SR-ROR (,39 mot ,25 i Utvalgsprøven 15).

Liksom på 5. trinn har bedömarna på 8. trinn bedömt mer likt varandra i Pilot 15 än Utvalgspröven 15. I den förra mätningen återfinns färre distinkta klasser av stränghet med *H* = 4,70 och *R* = ,91 mot Utvalgsprøven *H* = 8,92 och *R* = 98. På 8. trinn är dock andelen exakt överensstämmande bedömningar något färre i Pilot 15 (42,6 % mot 45,8), även om korrelationen är högre i Pilot 15 (*SR-ROR* = ,43) än Utvalgsprøven 15 (*SR-ROR* = ,27).[4]

[4] I denna rapport presenterar vi inte listor med fit-statistik för alla bedömare. På 5. trinn uppvisar 3 av 33 bedömare signifikant infit över 1,2 och på 8. trinn uppvisar 8 bedömare av 32 signifikant infit över 1,2. En noggrann undersökning av detta kan avslöja svårbedömda texter och oväntat bedömarebeteende men är, så länge resultaten i stort är goda, en del av Skrivesenterets interna kvalitetsarbete.

ICC-analysen bidrar till att förstärka bilden av ökad samstämmighet. Resultaten, som presenteras i tabell 3.16, visar att panelmedlemmarna med traditionella mått ligger över eftersträvansvärda värden (≥ ,7). Samstämmigheten är också betydligt högre än tidigare år. Två undantag gäller dock för (1) skalan interpunktion (i tabellen kallad V6) för 5. trinn och (2) skalan stavning (V5) för 8. trinn. I kommande bedömningsvägledningar bör möjligen beskrivningen av dessa skalor ses över.

Tabell 3.16. ICC-analyser

	Snitt	V1	V2	V3	V4	V5	V6
5. trinn	,74	,78	,80	,71	,75	,74	,63
8. trinn	,74	,80	,70	,80	,78	,61	,77

Notera. V = bedömningsområde. ICC = Intra Class Correlation Coefficient, average measure.

Det sista kvalitetsmåttet som presenteras rör den generaliserade skalan, det vill säga den skalstruktur som uppstår när vi utgår från elevernas samlade prestationer över alla bedömningsskalor (se Engelhard, 2013). Som nämndes tidigare finns speciella mått för kvaliteten på skalstrulturen och vi skall se närmare på *Rasch-Andrich Thresholds* (RAT). RAT anger var på logit-skalan som två kategorier är lika sannolika. Grafiskt illustreras detta i panelerna till vänster i figur 3.2. När bedömningsinstrumentet fungerar väl gäller följande: varje skalsteg är någon gång det mest sannolika och RAT-värdet bör avancera med minst 1,4 logits, men inte med mer än 5,0 logits. I tabell 3.17 återges resultaten för 5. trinn respektive 8. trinn.

Tabell 3.17. RAT för respektive trinn.

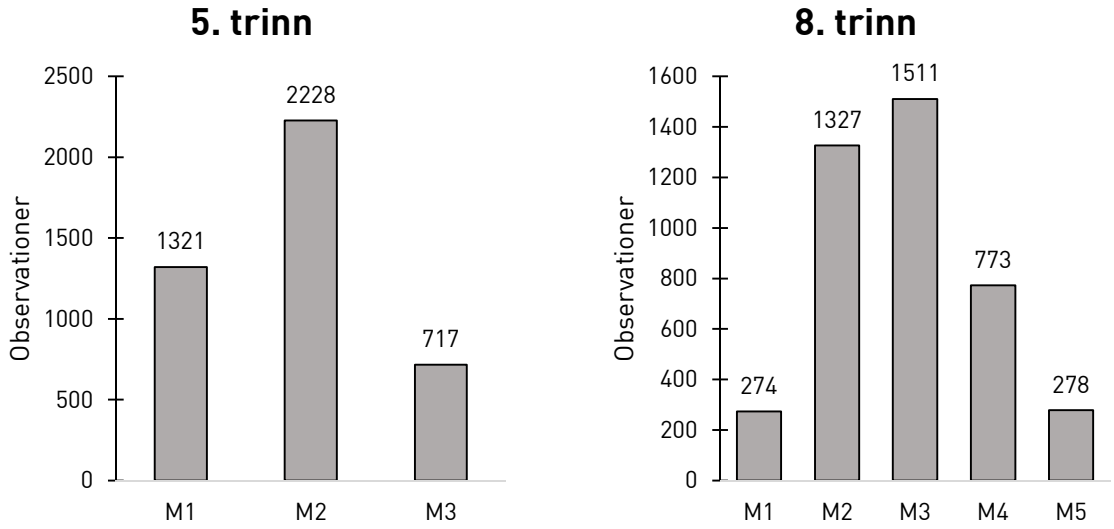
	5. trinn		8. trinn	
Skalsteg	Rasch-Andrich Thresholds	S.E.	Rasch-Andrich Thresholds	S.E.
1	–	–	–	–
2	-1,88	0,04	-3,58	0,08
3	1,88	0,05	-0,69	0,04
4	n/a	n/a	1,30	0,05
5	n/a	n/a	2,98	0,08

Notera. S.E. = mätfelet för Rasch-Andrich Thresholds.

Siffrorna bekräftar det intryck som graferna i figur 3.2 ger vid handen, nämligen att skalorna fungerar väl, när vi tar RAT i betraktande. Ett mått som också är av betydelse är fit-statistiken; denna bör indikera 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. För ingen av kategorierna är detta fallet.

Vi skall avslutningsvis se närmare på distributionen av utdelade poäng. Det ligger ingen förväntning om viss distribution bakom utvecklingen av Skriveprøven, men stark skevhet, för att ta ett exempel, kan indikera att vissa skalsteg inte fungerar optimalt. Om så är fallet på den generella skalan kan distributionen på de enskilda skalorna studeras närmare för att få en solid uppfattning om vari eventuella problem består.

I figur 3.3 återges distributionen för 5. respektive 8. trinn. För 5. trinn gäller att en knapp tredjedel (29 %) av observationerna är i kategori 1, medan drygt hälften av alla observationer är i kategori 2 (54 %). 17 % av observationerna är i kategori 3, som alltså är den minst använda. Även för 8. trinn ser vi att skalstegen i början av skalan flitigt använda, medan högre skalsteg används mer sällan. Kategori 1 och 2 har tillsammans 38 % av observationerna, medan kategori 3 ensamt svarar för 36 %. De två översta kategorierna har färre observationer (19 % respektive 6 %). Mönstret liknar det som tidigare rapporterats och givet att alla skalsteg åtminstone används i viss utsträckning finns ingen grund att utesluta något steg ur analysen.



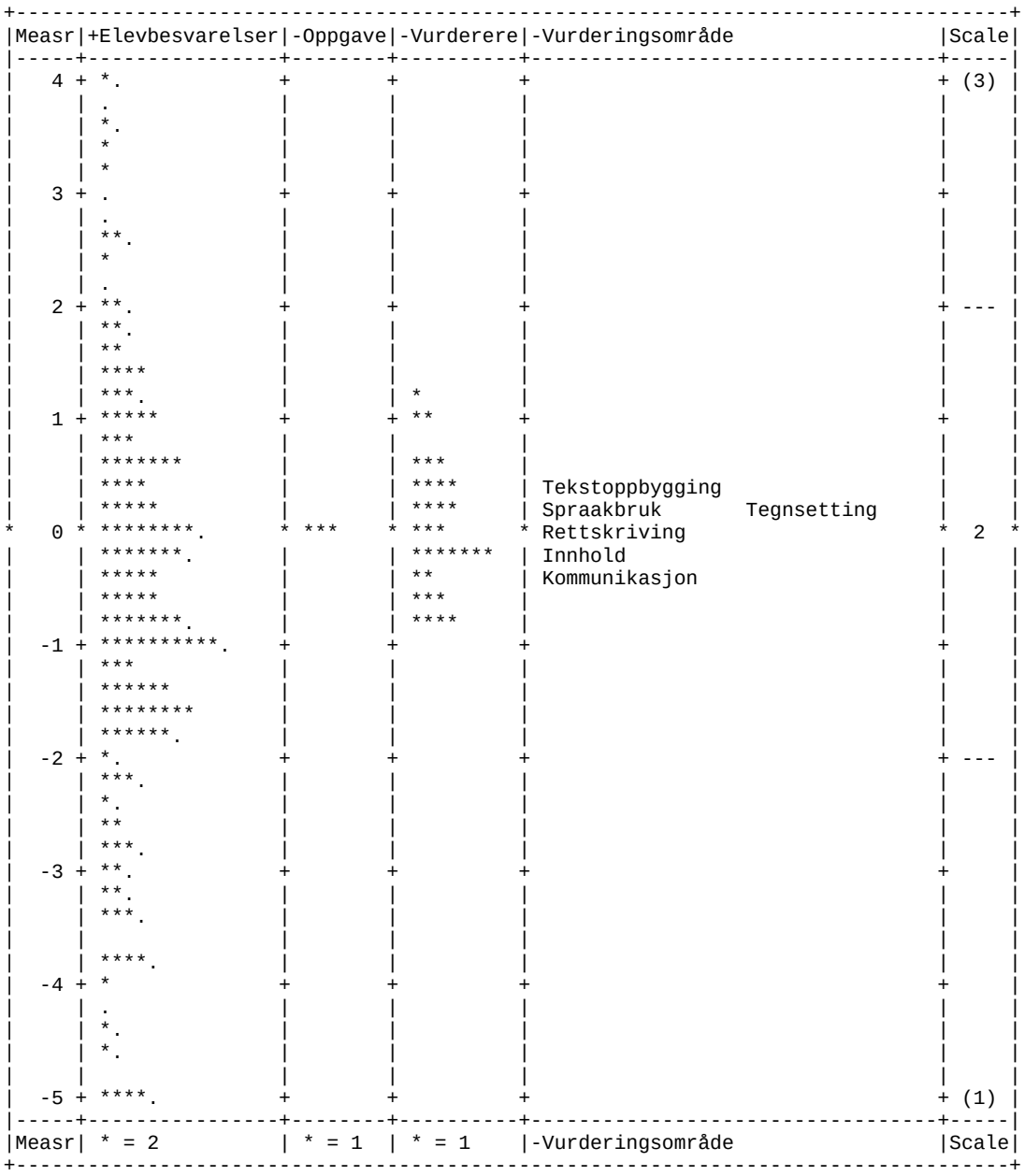
Figur 3.3. Distribution av poäng til 5. respektive 8. trinn.

Sammanfattningsvis visar resultaten att mätningen av elevförmåga är trovärdig och att skillnader mellan elever inte beror på slumpen. Detta är en viktig förutsättning för att kunna behandla de resultat som presenteras nedan som meningsfulla.

3.3.3 Resultaten

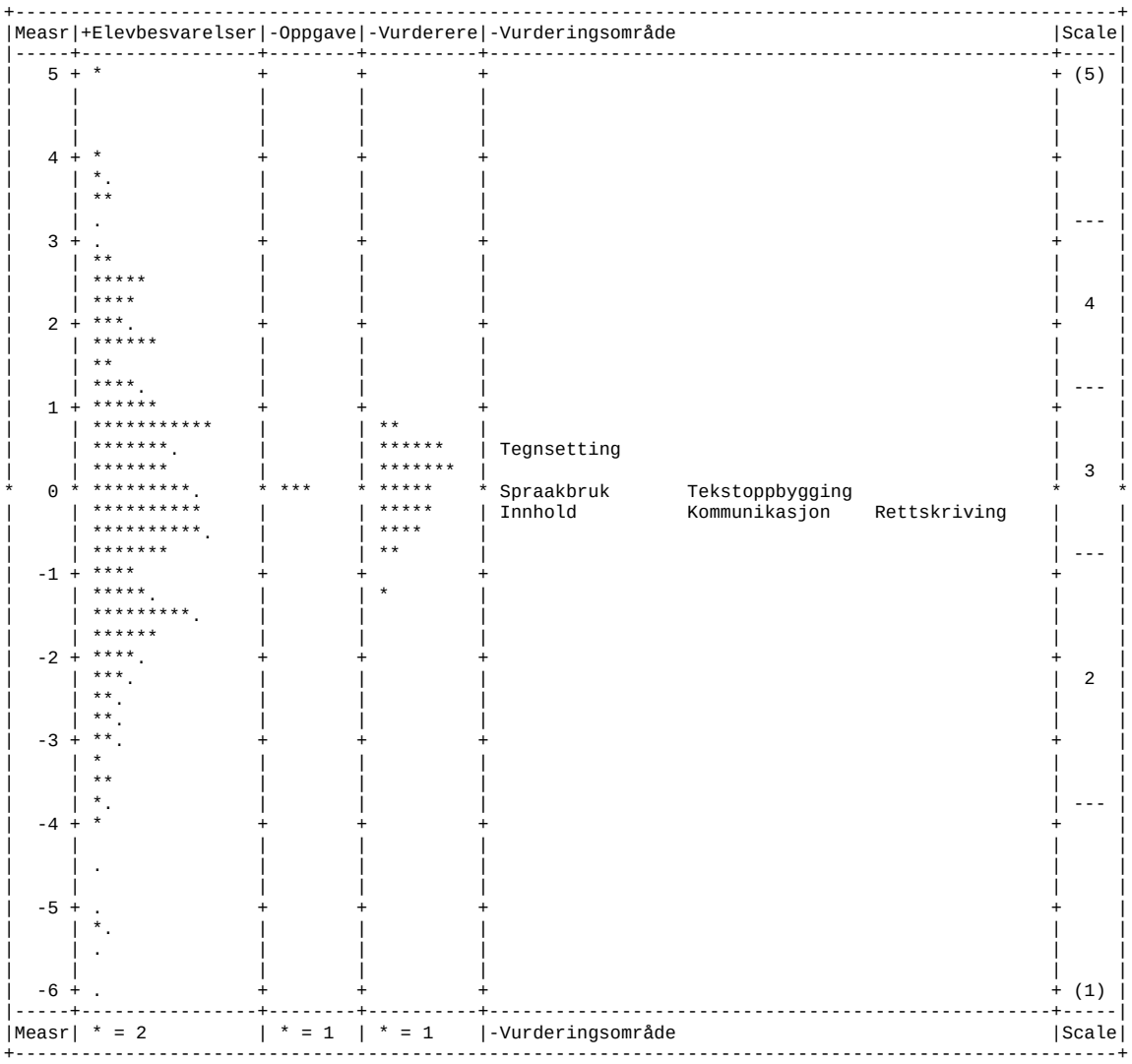
Resultaten presenteras i två format. Dels åskådliggörs dem med så kallade variabelkartor (också kallade t.ex. logit-kartor och Wright maps). Dessa kan läsas som histogram med logitskalan placerad till vänster och därefter olika facets som placerats på den skalan. Dels återges resultaten på *fair average-skalan*. Fair average transformerar logit-värdet tillbaka till ursprungsskalan och är på så vis enklare att tolka än logit-värden (som ofta går mellan -5 och 5).

Figur 3.4 återger resultatet av MFRM-analysen på 5. trinn. Figuren visar en del av det som tidigare berörts, nämligen att det är spridning bland personer (elever), bedömare och bedömningsområden. De flesta elever har ett förväntat resultat mellan 1,5 och 2,5, vilket också antyds i figur 3.3. Variabel-kartan visar dessutom relationen mellan de olika bedömningsområdena. I Pilot 15 var till exempel textstruktur svårast, medan kommunikation var enklast.



Figur 3.4. Variabel-karta för 5. trinn

Figur 3.5 återger resultatet av MFRM-analysen på 8. trinn. Också denna figur visar en del av det som tidigare berörts, nämligen att det är spridning bland personer (elever), bedömare och bedömningsområden. De flesta elever har ett förväntat resultat mellan strax under 2,5 och strax över 3,5, vilket också antyds i figur 3.3. Variabelkartan visar dessutom relationen mellan de olika bedömningsområdena. I Pilot 15 var till exempel interpunktion svårast, medan innehåll, kommunikation och stavning var enklast.



Figur 3.5. Variabel-karta för 8. trinn

I tabell 3.18 återges 5. trinnsresultaten på fair average-skalan för samlat resultat ("tot."), samt resultat på funktionskompetenser och resultat på kodningskompetenser. Resultaten presenteras för elevgruppen som helhelt och uppdelat på kön.[5]

[5] Att rapportera resultat på kön motiveras av två skäl. För det första finns det inte utan vidare skäl att anta att flickor och pojkar presterar olikt. Om de emellertid gör det är det intressant att studera hur den skillnaden ser ut mellan olika uppgifter. Avvikelser från mönster kan indikera att en uppgift gynnar eller missgynnar något av könen. Ett andra skäl är regeringens satsning *Språkløyper*, som har pojkar som särskild satsningsgrupp, men som saknar data på skillnaden mellan flickors och pojkars skrivande.

Tabell 3.18. Fair Average för 5. trinn.

	Tot.		Funksjon		Kode	
	M	S.D.	M	S.D.	M	S.D.
Total (*N* = 300[a])	1,84	0,46	1,85	0,52	1,82	0,47
Jente (*n* = 154)	1,97	0,44	2,01	0,50	1,92	0,47
Gutt (*n* = 146)	1,69	0,43	1,69	0,49	1,72	0,45

Notera. M = medel, S.D. = standardavvikelse. Funksjon = kommunikation, innehåll, textstruktur och språkbruk. Kode = stavning och interpunktion. [a]För en elev saknas uppgifter om kön. Denna elev är därför exkluderad ur sammanställningen.

Som vi kan se presterar elevgruppen som helhet strax under mittnivån på skalan, både samlat sett och uppdelat på funktions- och kodningskompetenser. Granskar vi resultaten närmare ser vi att flickorna presterar mycket nära mittnivån samlat sett, över mittnivån på funktionskompetenser och strax under på kodningskompetenser. Detta gäller inte för pojkar som jämt över får resultat som ligger under mittnivån. Resultaten ger därmed uttryck för stabilitet i jämförelse med tidigare mätningar.

I tabell 3.19 redovisas resultaten uppdelat på uppgifter. De elever som genomfört P1553 får högst resultat (1,88 i genomsnitt). På denna uppgift får flickorna resultat över mittnivån såväl samlat sett som uppdelat i funktions- och kodningskompetenser. Det får inte pojkarna, som jämt över får resultat under mittnivån. De elever som löst P1665a får i genomsnitt 1,83, vilket är marginellt högre än resultatet för de elever som löst P1566b (1,79). På båda overbevise-uppgifterna får flickorna högre resultat. Resultaten liknar dessutom varandra (1,94 respektive 1,95). De pojkar som skrivit P1556a får högre resultat än de pojkar som skrivit P1556b. Den senare uppgiften står ut i så måtto att de pojkar som skrivit denna uppgift får lägre resultat också än de pojkar som skrivit P1553.

Tabell 3.19. Fair Average uppgiftsnivå för 5. trinn.

		Tot.		Funksjon		Kode	
		M	S.D.	M	S.D.	M	S.D.
P1553	T (*N* = 101)	1,88	0,50	1,88	0,57	1,89	0,47
	J (*n* = 52)	2,04	0,48	2,05	0,54	2,00	0,47
	G (*n* = 49)	1,72	0,47	1,70	0,55	1,77	0,43
P1556a	T (*N* = 101)	1,83	0,42	1,86	0,46	1,79	0,46
	J (*n* = 53)	1,94	0,40	1,98	0,45	1,87	0,48
	G (*n* = 48)	1,72	0,40	1,73	0,44	1,70	0,42
P1556b	T (*N* = 98)	1,79	0,46	1,81	0,53	1,78	0,47
	J (*n* = 49)	1,95	0,44	1,99	0,51	1,87	0,44
	G (*n* = 49)	1,64	0,43	1,62	0,49	1,69	0,48

Notera. T = total, J = jente, G = gutt, M = medel, S.D. = standardavvikelse. Funksjon = kommunikation, innehåll, textstruktur och språkbruk. Kode = stavning och interpunktion.

Avslutningsvis har vi för 5. trinn beräknat skillnaden mellan flickor och pojkar, i termer av effektstorlek (Cohens *d*), samlat sett och uppdelat på uppgift. Till grund för beräkningarna ligger det samlade resultatet. I tabell 3.20 kan vi se att skillnaden va-

rierar mellan *d* = 0,55 för P1556a och *d* = 0,69 för P1556b. En effektstorlek på 0,55 innebär i det här fallet att omkring 71 % av pojkarna får lägre resultat än genomsnittsflickan, medan en effektstorlek på 0,69 innebär att omkring 76 % av pojkarna får lägre resultat än genomsnittsflickan (se bilaga 3). Resultaten indikerar P1556b bidrar till att öka skillnaderna mellan flickor och pojkar.

Tabell 3.20. Resultatskillnad mellan flickor och pojkar, 5. trinn.

	M_1	$S.D._1$	M_2	$S.D._2$	*t*	*df*	*p*	*ES*
Alla uppg.	1,97	0,44	1,69	0,43	5,59	298	0,00	0,64
P1553	2,04	0,48	1,72	0,47	3,39	99	0,00	0,67
P1556a	1,94	0,40	1,72	0,40	2,79	99	0,01	0,55
P1556b	1,95	0,44	1,64	0,43	3,43	96	0,00	0,69

Notera. M_1= Medel för jenter, $S.D._1$ = standardavvikelse för jenter, M_2 = medel för gutter, $S.D._2$ = standardavvikelse för gutter, ES = Cohens *d*.

När vi tar samtliga resultat i betraktande ser vi att resultaten i genomsnitt påminner om de som varit uppmätta förut, nämligen att elever presterar strax under skalans mittnivå. Vi ser också att P1556b är den uppgift där eleverna får lägst resultat och där skillnaden mellan flickor och pojkar är som störst. Detta är också den 5. trinnsuppgift som får lägst omdöme av lärarinformanterna.

På 8. trinn presterar eleverna i genomsnitt strax under mittnivån (2,87). Detta gäller dock inte för flickor som i genomsnitt presterar över mittnivån samlat sett (3,10) och uppdelat på funktions- och kodningskompetenser (3,15 respektive 3,01). Pojkarna presterar en bit under mittnivån samlat sett (2,67), vilket också gäller för funktions- och kodningskompetenser (2,71 respektive 2,61). Resultaten presenteras i tabell 3.20.

Tabell 3.20. Fair Average för 8. trinn.

	Tot.		Funksjon		Kode	
	M	S.D.	M	S.D.	M	S.D.
Total (N = 303)	2,87	0,80	2,91	0,86	2,79	0,83
Jente (n = 140)	3,10	0,78	3,15	0,83	3,01	0,82
Gutt (n = 163)	2,67	0,77	2,71	0,84	2,61	0,79

Notera. M = medel, S.D. = standardavvikelse. Funksjon = kommunikation, innehåll, textstruktur och språkbruk. Kode = stavning och interpunktion.

När vi ställer upp resultaten uppdelat på uppgift ser vi att de elever som löst P1586 får högst resultat med 3,08 i genomsnitt för hela gruppen, medan de som löst P1583 får lägst resultat (2,75). Resultatet på den förra är marginellt lägre än resultatet för P1584 (2,78). Studerar vi skillnaden mellan flickor och pojkar kan vi konstatera att den också på 8. trinn är konstant, men att en uppgift, P1583, står ut. Här får flickorna i genomsnitt 3,15 medan pojkarna får 2,39.

Tabell 3.21. Fair Average uppgiftsnivå för 8. trinn.

		Tot.		Funksjon		Kode	
		M	S.D.	M	S.D.	M	S.D.
P1583	T (N = 101)	2,75	0,86	2,76	0,93	2,73	0,86
	J (n = 48)	3,15	0,78	3,20	0,85	3,04	0,75
	G (n = 53)	2,39	0,78	2,37	0,81	2,45	0,86
P1584	T (N = 101)	2,78	0,80	2,78	0,83	2,78	0,87
	J (n = 45)	2,93	0,82	2,93	0,84	2,93	0,91
	G (n = 56)	2,65	0,77	2,66	0,81	2,65	0,81
P1586	T (N = 101)	3,08	0,71	3,19	0,76	2,87	0,75
	J (n = 47)	3,22	0,74	3,30	0,77	3,05	0,81
	G (n = 54)	2,97	0,67	3,10	0,74	2,72	0,67

Notera. T = total, J = jente, G = gutt, M = medel, S.D. = standardavvikelse. Funksjon = kommunikation, innehåll, textstruktur och språkbruk. Kode = stavning och interpunktion.

Även på 8. trinn har vi undersökt skillnaden mellan flickor och pojkar i termer av effektstorlek. Den samlade skillnaden är, liksom alla skillnader på 5. trinn, signifikant och motsvarar här *d* = 0,55. Annorlunda uttryckt betyder det cirka 72 % av pojkarna får lägre resultat än genomsnittsflickan. När vi delar upp resultaten på uppgiftsnivå ser vi emellertid att skillnaderna är små och icke-signifikanta för P1584 och P1586 (*d* = 0,35 respektive *d* = 0,34), medan signifikant och relativt stor för P1583 med *d* = 0,96. På denna uppgift presterar alltså cirka 83 % av pojkarna sämre än genomsnittsflickan.

Tabell 3.20. Resultatskillnad mellan flickor och pojkar, 8. trinn.

	M_1	$S.D._1$	M_2	$S.D._2$	*t*	*df*	*p*	*ES*
Alla uppg.	3,10	0,78	2,67	0,77	4,77	301	0,00	0,55
P1583	3,15	0,78	2,39	0,78	4,85	99	0,00	0,96
P1584	2,93	0,82	2,65	0,77	1,76	92	0,08	0,35
P1586	3,22	0,74	2,97	0,67	1,74	99	0,09	0,34

Notera. M_1= Medel för jenter, $S.D._1$ = standardavvikelse för jenter, M_2 = medel för gutter, $S.D._2$ = standardavvikelse för gutter, ES = Cohens *d*.

När vi tar samtliga resultat i betraktande gäller även för 8. trinn att resultaten i genomsnitt påminner om de som varit uppmätta förut, nämligen att elever presterar strax under skalans mittnivå. På 8. trinn är emellertid nyanserna större. På uppgift P1586 presterar hela elevgruppen i genomsnitt över mittnivån och uppgift P1583 är den som ensamt bidrar till signifikant skillnad mellan könen. Här är skillnaden så stor att det är rimligt att misstänka att pojkarna underpresterar. Några indikationer på det har emellertid inte förekommit i svaren från lärare eller panel.

4 Sammanfattning

Utgångspunkten för denna rapport har varit frågan: är det rimligt att anta att uppgifterna kan fungera i utvalgsprøven? Rapporten har redovisat resultat av tre undersökningar. För det första har medverkande lärare tillfrågats om synpunkter på aspekter av hur väl uppgifterna fungerar. Också medlemmar i bedömningspanelen har lämnat sina synpunkter. För det andra har Skrivesenteret observerat genomförande av Pilot 15 för att bilda sig en uppfattning om i vilken utsträckning instruktioner och annat material möjliggör för lärare att agera på ett sådant sätt som Skrivesenteret avsett. För det tredje har drygt 600 elevsvar bedömts av bedömningspanelen. Dessa bedömningar har underkastats faktoranalyser, CTT-analyser och MFRM-analyser.

Resultaten visar att fyra av sex uppgifter överlag får godkänt betyg av lärarna. Det gäller P1553, P1556b, P1583 och P1586. De resterande två uppgifterna, P1556b och P1584, möter däremot viss kritik. Allvarligast är invändningarna mot påståendet att uppgifterna ger möjlighet för eleverna att visa sina skrivfärdigheter och mot påståendet att eleverna är motiverade. Uppgifter som inte uppfattas som goda eller där eleverna inte är motiverade kan ha svårt att bli del av populära Skriveprøver.

Bedömarpanelen är generellt positiva till uppgifterna. Ingen 5. trinnsuppgift tycks vara särskilt problematisk, men två uppgifter på 8. trinn uppfattas delvis som bristfälliga. Det gäller P1584 som mer än 20 % av panelmedlemmarna tycker är dålig och P1583, som lite mer än 10 % av medlemmarna tycker fungerar dåligt.

Observationerna av genomförandet visar att instruktioner och annat material fungerar väl. Alla lärare leder genomförandet på sådant sätt som Skrivesenteret avsett. Skrivesenteret noterar också viss variation i elevers engagemang och tidsbruk, men inget som kan relateras till bristfälliga instruktioner eller annat bristfälligt material från Skrivesenteret.

Analysen av genomförda bedömningar visar (a) indikationer på en en-faktormodell och (b) att data passar MFRM-modellen väl. Analysen visar också att samstämmigheten mellan bedömare för första gången i Skriveprøvens historia med god marginal ligger över den kritiska gränsen på ,70.

Själva resultaten pekar på att eleverna i genomsnitt presterar strax under skalans mittnivå. Variationerna är dock många. Genomsnittsresultaten på enskilda uppgifter varierar och det finns en i det närmaste konstant stor skillnad mellan flickor och pojkar. De två uppgifter där denna skillnad är som störst är P1556b, som kritiserats och, något förvånande, P1583, som hittills passerat någorlunda obemärkt.

Denna undersökning kan inte ge ett entydigt svar på frågeställning presenterad ovan. Den huvudsakliga orsaken är svaret varierar med uppgift.[6] Vår rekommendation är

[6] En annan, mycket viktig orsak är att varje elev bara skrivit en uppgift, vilket gör dem mindre jämförbara.

därför att vissa uppgifter blir kvar i systemet, medan de andra utgår. Vi föreslår att följande blir kvar: P1553, P1556a och P1586. Vi föreslår att följande uppgifter utgår till följd av kritik från lärare, panel eller avvikande prestationsmönster: P1556b, P1583 och P1584.

Under normala omständigheter skulle rekommendationerna medföra att Utvalgsprøven 2016 bara skulle inkludera tre uppgifter. Med den nya utvalsgprøve-designen är detta emellertid inte fallet.

Litteratur

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

Berge, K. L., Evensen, L. S., & Thygesen, R. (2016). The Wheel of Writing: a model of the writing domain for the teaching and assessing of writing as a key competency. *The Curriculum Journal*, 1–18. http://doi.org/10.1080/09585176.2015.1129980

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

Eckes, T. (2015). *Introduction to many-facet Rasch measurement* (2. utg.). Frankfurt am Main: Peter Lang.

Engelhard, G. (2013). *Invariant Measurement*. New York: Routledge.

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

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

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

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

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

Skar, G. B., & Iversen, J. M. (2016). *Nasjonale utvalgsprøver i skriving 2015. Teknisk rapport*. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.

Bilaga 1

Ingångstext, P1584

For 65 millioner år siden skjedde noe som utryddet over halvparten av alle artene som levde på jorda. Dinosaurene, som hadde levd på jorda i 160 millioner år, forsvant. Dette er en av naturhistoriens største gåter: hvorfor døde dinosaurene ut? Forskere har ulike teorier om hva som kan ha skjedd.

Det vi vet, er at det var stor aktivitet i vulkanene på jorda på den tiden. Gass fra vulkanutbrudd inneholdt mye svovel, og dette kan ha skapt sur nedbør. Røyk og aske fra vulkanutbrudd kan også ha skygget for sola slik at klimaet ble kaldere. Innholdet av et stoff som heter selen, økte i jordsmonnet. Selen er et stoff som er veldig giftig for fostre, og dette kan ha vært skadelig for eggene til dinosaurene.

Forskere har også funnet et stort krater som de mener kan stamme fra en stor asteroide som traff jorda. Dette skapte sannsynligvis store tidevannsbølger og store mengder støv som virvlet opp og blokkerte sollyset. Dette førte igjen til at plantene sluttet å vokse, og det ble stor mangel på mat.

Noen forskere mener også at insekter var bærere av dødelige sykdommer, og at disse bidro til at det brøt ut sykdomsepidemier blant de store dyrene. Insektene kan ha bidratt til spredning av plantesykdommer, og dinosaurene hadde dermed ikke like god tilgang til mat som før.

Forskerne er ikke helt sikre på hvorfor dinosaurene forsvant, og dette er fortsatt et uløst mysterium. Hva mener du kan ha skjedd?

Bilaga 2

Underlag för observation av genomförande av skrivprov

1. Är det möjligt för läraren att följa lärarinstruktionen? Om nej, vilka avsteg måste göras?
2. Vilka frågor ställer eleverna under genomgången?
3. Talar lärare eller elever om hur provresultatet kommer att användas?
4. Vad sker under «idémyldringen?» Är eleverna aktiva? Vilket spår för lärarna in eleverna på?
5. Verkar eleverna förstå vad de ska skriva om? Verkar eleverna förstå hur de skall producera texten?
6. Kommer eleverna igång med uppgiften snabbt eller långsamt?
7. Har eleverna frågor under tiden?
8. Är tiden för kort, för lång eller tillräcklig för merparten av eleverna?
9. Verkar eleverna motiverade? (Frågar t.ex. eleverna varför de skall genomföra provet?)
10. Vad har läraren för kommentarer eller frågor efter genomfört prov?

Bilaga 3

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