Evaluering af forsøgsprøver

i biologi, fysik/kemi og geografi

Martin Foldager Hindsholm, Hans Henrik Sievertsen, Marianne Mikkelsen,
Katrine Holm Filtenborg Kitchen og Morten Hjortskov Larsen

VIVE

Evaluering af forsøgsprøver
– i biologi, fysik/kemi og geografi

e-ISBN: 978-87-7582-309-3
Projekt: 302410
Finansiering: Styrelsen for Undervisning og Kvalitet

VIVE
Det Nationale Forsknings- og Analysecenter for Velfærd
Herluf Trolles Gade 11
1052 København K
www.vive.dk

VIVE støtter FN's verdensmål og angiver her, hvilket eller hvilke verdensmål der knytter sig til publikationen.

Forord

På baggrund af den nationale naturvidenskabsstrategi har Styrelsen for Undervisning og Kvalitet (STUK) opstartet et forsøg med grundskolens udtræksprøver i fysik/kemi, biologi og geografi. Forsøgsprøverne involverer blandt andet udvidelse af prøvernes varighed samt introduktionen af både simuleringsopgaver og skriftlige besvarelser. Til sammenligning er de eksisterende prøver selvrettende og består udelukkende af multiple choice-spørgsmål.

VIVE har på STUKs opdrag haft til opgave at evaluere forsøgsprøverne med henblik på at etablere grundlag for, at der kan træffes beslutning om ændringernes eventuelle permanentgørelse. Evalueringen er finansieret af STUK.

En stor tak skal lyde til alle de lærere, elever og censorer, som har bidraget til dataindsamlingen. Også en stor tak til både interne og eksterne kvalitetssikrere samt studentermedhjælperne, der har bidraget til både dataindsamling og analyse. Sidst, men ikke mindst, takker vi STUK for et rigtig godt, konstruktivt samarbejde om evalueringen.

Evalueringen er gennemført af projektleder og chefanalytiker Martin Foldager Hindsholm, Professor MSO Hans Henrik Sievertsen, senioranalytiker Marianne Mikkelsen, seniorforsker Morten Hjortskov Larsen samt universitetspraktikant Katrine Holm Filtenborg Kitchen.

Carsten Strømbæk Pedersen
Forsknings- og analysechef for VIVE Børn og Uddannelse

Indholdsfortegnelse

DEL 1
Afrapportering

Hovedresultater

Grundskolens naturfag får i disse år stor opmærksomhed, blandt andet gennem den naturvidenskabsstrategi, som den daværende regering udkom med i 2018. Den skitserer en samfundsmæssig virkelighed med mangel på naturvidenskabeligt uddannet arbejdskraft og et behov for en styrkelse af naturfagsundervisningen hele vejen gennem uddannelseskæden. Herudover involverer naturvidenskabsstrategien blandt andet et konkret ønske om at påbegynde en videreudvikling af prøverne i de tre naturfag i grundskolen: biologi, fysik/kemi og geografi. Konkret lægger strategien op til, at prøverne videreudvikles på forsøgsbasis, eksempelvis gennem udvidelse af prøvernes varighed og opgaveantal med det formål at afprøve eleverne bedre i både bredden og dybden. I forlængelse heraf igangsatte Børne- og Undervisningsministeriet et forsøg. Grundet covid-19-pandemien, der førte til aflysninger af udtrækprøverne i både 2021 og 2022, har det dog først i forbindelse med afgangsprøverne i sommeren 2023 været muligt for første gang at afprøve de tilpassede forsøgsprøver.

VIVE har på Styrelsen for Undervisning og Kvalitets opdrag haft til opgave at følge forsøget med det formål at evaluere forsøgsprøvernes kvalitet og ændringernes betydning med henblik på at bidrage til at skabe grundlag for at beslutte, om forsøgsprøverne skal permanentgøres eller ej. Nedenfor beskriver vi først kort forsøgsprøverne, hvorefter evalueringens hovedresultater og konklusioner præsenteres. Disse udgør de fund, som vi har vurderet som de vigtigste, og som besvarer de væsentligste undersøgelsesspørgsmål i opdraget, men de er ikke udtømmende.

Forsøgsprøverne

De skriftlige udtrækprøver, der afprøves i forsøget, omtales som forsøgsprøver. De væsentligste nye elementer, der indgår i forsøgsprøverne, er 1) anvendelse af simuleringer og 2) mulighed for at angive korte tekstsvar.

Fakta om undersøgelsen

Evalueringen baserer sig på et omfattende datagrundlag, der både inkluderer prøvedata, registerdata, interview af lærere og elever samt en spørgeskemaundersøgelse blandt lærere.

Forsøgsprøverne består af to dele – del A og del B. Del A svarer til den eksisterende prøve, der består af selvrettende multiple choice-spørgsmål, dog med færre opgaver (trefjerdedele af den eksisterende prøve). Del B udgør den nye del af prøverne og består af opgaver, hvor eleverne skal formulere korte, åbne tekstsvar. Ved disse opgaver kan der indgå na-

turfaglige simuleringer, video og animationer. Simuleringer er et interaktivt element i prøven som fx en undersøgelse fra fysik/kemi, hvor eleven kan titrere væsker og aflæse resultater og dermed forholde sig til egen empiri.

Sammenlagt har forsøgsprøverne en tidsramme på 90 minutter, hvor de nuværende udtræksprøver har en tidsramme på 60 minutter. Eleverne styrer selv, hvor lang tid de bruger på hver del af forsøgsprøverne, men de er designet, så det forventede tidsforbrug er 45 minutter på hver del. Del A af forsøgsprøverne gennemført i sommeren 2023 bestod af mellem 54 og 58 delopgaver, mens del B bestod af mellem 17 og 18 delopgaver (herefter omtalt 'items'). Alle elever får de samme opgaver i samme rækkefølge, og hvert år formuleres der nye opgaver af en opgavekommission for hvert fag.

Forsøgsprøverne afprøver ligesom de nuværende udtræksprøver eleverne i kompetencemålene og de faglige områder inden for færdigheds- og vidensområderne for hhv. biologi, geografi og fysik/kemi, som de er beskrevet i 'Fælles Mål'. Både de eksisterende prøver og forsøgsprøverne gennemføres i udgangspunktet digitalt ved en computer, medmindre prøven aflægges på særlige vilkår.

Da forsøgsprøverne indeholder opgaver, der skal besvares med elevers selvstændige formuleringer, bliver prøvens del B bedømt af et statsligt beskikket censorkorps. Eleverne får en samlet karakter for del A og del B (https://www.uvm.dk/folkeskolen/folkeskolens-proever/proevetilrettelaeggelse/adgang-tilmelding-og-booking/proeveformer-og-forsoeg/forsoeg-med-de-skriftlige-proever-i-naturfag/generel-information-om-forsoeget).

125 folkeskoler og frie grundskoler deltog i forsøget i skoleåret 2022/2023. På 83 skoler blev mindst én klasse udtrukket til en af prøverne.

Prøvernes sværhedsgrad passer til elevernes dygtighed, og den nye del supplerer den eksisterende del godt

Det er essentielt, at prøver har en passende sværhedsgrad for effektivt at kunne vurdere elevernes dygtighed. Vi finder, at den nye del af prøverne, del B, generelt er lidt sværere end den eksisterende del, del A, som vi bruger som sammenligningsgrundlag. Sværhedsgraden ligger dog i alle tilfælde på et fornuftigt niveau, hvor prøven hverken er for svær eller for nem. Herudover supplerer del A og del B hinanden godt. Mens del A er god til at identificere eleverne rundt om gennemsnittet, er del B bedre egnet til at identificere de meget dygtige elever og de mindre dygtige elever. Herudover viser analyserne, at biologiprøven er den nemmeste af de tre prøver, og at den nemmeste opgavetype er simuleringsopgaverne, hvor eleverne skal følge anvisningerne på skærmen og aflæse et svar. Det vil sige, at der gennemsnitligt er en større an-

del af eleverne, der svarer rigtigt på opgaverne i biologiprøven, og at de opgaver, som flest elever svarer rigtigt på – på tværs af prøverne – er simuleringsopgaverne.

De grundlæggende prøveegenskaber viser visse tegn på udfordringer, men ikke i særlig grad på grund af ændringerne

Et grundlæggende formål med grundskolens afgangsprøver er at give et validt og pålideligt mål af elevernes dygtighed i det givne fag. Validitet og pålidelighed betyder i den sammenhæng, at en prøve giver ensartede konklusioner, hvis den gentages, og at den måler en entydig underliggende dimension for dygtighed i faget. Forskningsfeltet for prøver og test kaldes testteori, og i testteorien har man udviklet en række rammer og modeller for prøver og test. En af de mest anvendte modeller er Rasch-modellen. Hvis en prøve opfylder Rasch-modellens antagelser, giver prøven valide og pålidelige mål for elevernes dygtighed.

Med udgangspunkt i Rasch-modellen har vi undersøgt fire grundlæggende egenskaber i naturfagsprøverne: intern konsistens, lokal uafhængighed, endimensionalitet og stabilitet. Vi har undersøgt egenskaberne både for del A og del B, hvor del A svarer til den eksisterende prøve og dermed udgør et benchmark for den nye del, del B. På tværs af alle egenskaber ser vi brud med Rasch-modellens antagelser i både del A og del B. Vi ser dog ikke nogen tegn på, at udfordringerne er større i den nye del, del B, sammenlignet med den eksisterende del, del A.

I alle tre fag er den **interne konsistens** acceptabel eller god for almindelige test. Men for såkaldte high-stakes-test, som afgangsprøverne, vil man ofte forvente en lidt højere grad af konsistens. Konsistens betyder, at der er sammenhæng i prøvernes delspørgsmål, og at de måler samme underliggende dygtighed. En høj grad af konsistens er særligt vigtig i situationer, hvor eleverne kun besvarer en delmængde af prøvernes spørgsmål, fordi det så kan betyde, at to elever med samme målte dygtighed i virkeligheden er dygtige på vidt forskellige områder og ikke har samme dygtighed i samme fag.

Lokal uafhængighed betyder, at alle spørgsmål i prøven bidrager med uafhængig ny viden om elevernes dygtighed. Hvis to spørgsmål er afhængige, så vil en elevs antal point for en besvarelse afhænge af elevens besvarelse af et andet spørgsmål. Det vil give et misvisende billede af elevens dygtighed og samtidig være en ineffektiv måde at afprøve eleverne på. Vi finder brud på antagelsen om lokal uafhængighed i alle tre fag og i både del A og del B, men det er primært i del B i fysik/kemi og geografi, at vi finder en lidt højere andel af spørgsmålspar, hvor der er en sammenhæng i besvarelserne.

Endimensionalitet betyder, at prøverne kun måler dygtigheden på én underliggende dimension. Konkret kan man for eksempel forestille sig, at fysik/kemi-prøven afprøver en dimension af dygtighed, der gælder fysik, og en anden dimension, der gælder kemi. Hvis de to dimensioner ikke er konsistente, således at elever, der er dygtige i fysik, også er dygtige i kemi, så kan det betyde, at to elever med samme resultat i prøverne i virkeligheden er dygtige i to forskellige fag. Også for endimensionalitet finder vi brud i alle tre fag og i begge dele, men udfordringerne ser ud til at være lidt større i del A end i del B.

Stabilitet dækker over, om prøvens spørgsmål giver stabile svar, således at for eksempel et svært spørgsmål i en prøve primært besvares rigtigt af dygtige elever. En lav grad af stabilitet betyder, at det er mere eller mindre tilfældigt, om eleverne svarer rigtigt eller forkert. I undersøgelsen af spørgsmålenes stabilitet finder vi også, at andelen af spørgsmål med udfordringer er større i del A i fysik/kemi og geografi, mens udfordringerne er lidt større i del B i biologi.

Anvender vi del A, der svarer til de eksisterende prøver, som benchmark, er der altså ikke tegn på, at forsøgsprøverne repræsenteret ved del B er dårligere konstrueret. De er dog heller ikke bedre konstrueret.

Det er særligt vigtigt, at prøver, som gentages på flere kohorter og elevgrupper, og prøver, hvor eleverne udtrækker en række opgaver fra en opgavebank, opfylder Rasch-modellens antagelser, fordi det betyder, at målingerne er sammenlignelige på tværs af kohorter og på tværs af de opgaver, eleverne gives. Når alle elever modtager alle opgaver, og prøverne formuleres og designes specifikt til hver kohorte – som i dette tilfælde – er implikationerne mindre problematiske, men det betyder ikke, at man ikke bør vurdere, om prøvedesignet kan forbedres. Rasch-modellens antagelser er forholdsvis restriktive, og et brud på dem skal derfor afvejes relativt til andre mål såsom at afprøve dygtighederne på forholdsvis bredt definerede områder og med varierede metoder.

Danskkundskaber har ikke større betydning på trods af introduktionen af skriftlige svar

Både lærere og opdragsgiver har udtrykt bekymring for, om tilføjelsen af skriftlige svar til prøverne vil favorisere de sprogligt stærke elever og omvendt stille de sprogligt udfordrede elever ekstra dårligt. Det finder vi dog ingen tegn på. Generelt tyder vores analyser således på, at ændringerne *ikke* stiller bestemte elevgrupper dårligere end ved de eksisterende prøver, og at der ikke er grund til bekymring. Konkret finder vi, at elevernes danskfærdigheder ikke har større betydning for, hvordan de klarer sig i forsøgsprøvernes del B sammenlignet med del A. Tilsvarende er der heller ikke tegn på, at baggrundsfaktorer som biologisk køn, oprindelse og forældres uddannelse har større betydning i del B sammenlignet med del A.

Der er god enighed mellem censorerne

Mens de eksisterende prøver er selvrettende, kræver den nye del af forsøgsprøverne en censorbedømmelse. Det skal helst ikke være sådan, at en elevs resultat afhænger af, om eleven er heldig eller uheldig med at få en gavmild eller mindre gavmild censor. I stedet bør det være sådan, at to bedømmere af samme besvarelse typisk giver samme resultat. Alle gennemførte forsøgsprøver er vurderet af to tilfældige og uafhængige censorer, og vi har derfor kunnet undersøge deres enighed. Vi finder, at pålideligheden på tværs af censorbesvarelser – eller enigheden mellem censorerne – er acceptabel i fysik/kemi og fremragende i geografi og biologi.

Lærere og elever: Et skridt i den rigtige retning, men skriftligheden opleves svær

Lærerne hilser generelt prøveændringerne velkomne. Både lærere og elever sætter pris på den større variation i opgavetyper, og særligt introduktionen af simuleringer fremhæves som noget positivt. Det kommer blandt andet til udtryk ved, at 77 % af lærerne mener, at forsøgsprøverne overordnet set er et skridt i den rigtige retning. Et stort flertal vurderer ligeledes, at forsøgsprøverne afprøver elevernes naturfaglige kompetencer, og lærerne er generelt af den holdning, at forsøgsprøverne er mere nutidige og bedre matcher undervisningen og de faglige mål. Men skriftligheden opleves svær, og det kan være uklart for eleverne, hvad et godt skriftligt svar indebærer. At skriftligheden opleves svær, forklares dog også med, at den kræver, at eleverne i højere grad bringer deres viden og kompetencer i spil og anvender relevante naturfaglige begreber. De skal kunne argumentere og beskrive, og det betyder, at eventuelle huller i deres viden og kompetencer ikke lige så nemt kan skjules. I del A har eleverne via multiple choice-formatet altid 25 % chance for at svare rigtigt (ved fire svarmuligheder), selvom de ikke kender svaret. Til sammenligning vil de have meget svært ved at gætte sig til et rigtigt svar i del B.

Der arbejdes mere med simuleringer og skriftlighed, og opmærksomheden på elevernes naturfaglige kompetencer er øget

Mens tegnene på ændringer i lærernes undervisning så meget begrænsede ud i starten af forsøget, ser ændringerne nu ud til at have taget fart. Særligt har lærerne i højere grad implementeret simuleringer i deres undervisning. Simuleringer er blandt andet en vej til at arbejde med elevernes modelleringskompetence og nævnes i fagenes læseplaner som et eksempel på en interaktiv model. Skriftligheden fylder også mere, og 54 % af lærerne svarer, at de som følge af prøveændringen har været mere opmærksomme på at udvikle elevernes naturfaglige kompetencer.

Det skriftlige arbejde har fokus på argumentation, forklaring og fagbegreber og øves oftest i forbindelse med forsøg og undersøgelser.

Lærerne ser gode muligheder i brugen af simuleringer og anvender dem især til at visualisere abstrakte fænomener og processer, men også som supplement til fysiske forsøg eller i tilfælde, hvor undersøgelsen ellers ikke ville kunne gennemføres. Eleverne gives typisk stor frihed i arbejdet med simuleringer.

Lærerne fokuserer fortsat på at forberede eleverne til den fælles prøve, men balancen mellem prøverne vurderes styrket

De skriftlige naturfagsprøver er udtræksprøver, mens den fælles naturfagsprøve er obligatorisk og kræver gennemførsel af en række fællesfaglige forløb. Derfor er det kun naturligt, at lærerne primært fokuserer på at forberede eleverne til den fælles prøve i deres daglige undervisning. Og det har forsøgsprøverne ikke ændret. Lærernes fokus er fortsat på den fælles prøve. Først ved offentliggørelse af en klasses udtræk til en af de skriftlige prøver, får den et mere eksplicit fokus. Men i kraft af at lærerne oplever, at forsøgsprøverne i højere grad end de eksisterende prøver afprøver elevernes naturfaglige kompetencer, som også er fokus for den fælles prøve, vurderer lærerne, at forsøgsprøverne matcher den fælles prøve bedre, end de eksisterende udtræksprøver gør. Dermed vil undervisningen – som fokuserer på at forberede eleverne til den fælles prøve – formentlig også forberede eleverne bedre til forsøgsprøverne end til de eksisterende udtræksprøver. Den holdning er i hvert fald at spore hos lærerne. Direkte adspurgt svarer 38 % af lærerne, at de som følge af prøveændringerne har bedre mulighed for både at forberede eleverne til den fælles prøve og udtræksprøverne.

Der efterspørges mere konkret undervisningsmateriale

Lærerne udtrykker generelt begejstring over at tage del i forsøget, og stemningen omkring både forsøget og forsøgsprøverne er overvejende god. Enkelte lærere oplever, at kommunikationen omkring forsøget kan styrkes, men den barriere for forandring, der fylder klart mest blandt lærerne, er manglen på konkret undervisningsmateriale. Konkret efterspørges der i høj grad flere simuleringer og gerne på dansk. Foreløbig er mange læreres eneste kilde til simuleringer hjemmesiden PhET Colorado, som er en offentligt tilgængelig ressource, fra før forsøget blev startet op. Det er den eneste, de kender. Herudover efterspørger lærerne skriftlige opgaver og flere eksempler på, hvordan opgaverne i forsøgsprøverne kan se ud. 56 % af lærerne udtrykker i spørgeskemaundersøgelsen utilfredshed med tilgængeligheden af øvelsesopgaver og simuleringer til undervisningen.

Konklusion

Samlet set viser evalueringen, at mens forsøgsprøverne opfylder visse centrale kvalitetskrav, er der potentielt plads til forbedring, især med hensyn til at sikre større stabilitet og uafhængighed mellem opgaverne. Det er vigtigt at fortsætte med at overvåge og justere prøverne for at forbedre deres pålidelighed og validitet, særligt i lyset af de udfordringer, der er identificeret i forhold til endimensionalitet, lokal uafhængighed og stabilitet. Dette vil hjælpe med at sikre, at prøverne forbliver et retfærdigt og nøjagtigt mål for elevernes dygtighed.

Udfordringerne med Rasch-modellens grundantagelser er dog ikke større i prøvernes nye del sammenlignet med den eksisterende del, og vores analyser indikerer derfor ikke, at tilføjelsen af blandt andet korte tekstsvar og interaktive simuleringer gør prøveegenskaberne dårligere.

Vi finder heller ikke problemer i, at ændringerne medfører behov for manuelle vurderinger af prøvebesvarelserne. Censorerne er gode til at følge rettevejledningerne, og der er acceptabel enighed mellem dem. Samtidig viser vores øvrige analyser, at både lærere og elever grundlæggende ser positivt på ændringerne, og at lærerne mener, at man med forsøgsprøverne er gået et skridt i den rigtige retning. Der er bekymringer omkring, om skriftligheden i forsøgsprøverne vil have konsekvenser for de sprogligt udfordrede elever og stille dem dårligere end ved de eksisterende prøver, men vi finder ingen tegn på, at det faktisk er tilfældet i de statistiske analyser. Herudover er der dog blandt lærerne en ret udbredt efterspørgsel på, at de understøttes bedre i at lave de nødvendige ændringer af deres undervisning – konkret gennem mere undervisningsmateriale – men lærerne er allerede godt i gang på trods af en oplevelse af mangel på tid.

Alt i alt er vores konklusion på den baggrund, at der hverken er noget ved selve ændringerne i forsøgsprøverne eller i oplevelsen af dem og deres konsekvenser, der står i vejen for, at ændringerne gøres permanente. Det er dog væsentligt fortsat at overvåge og justere prøverne, arbejde med kommunikationen til skolerne og at sikre, at lærerne støttes tilstrækkeligt i implementeringsprocessen.

1 Indledning

I en tid med stigende teknologisering og store samfundsudfordringer med et naturvidenskabeligt islæt er der behov for børn og unge, der forstår og kan forholde sig til naturvidenskab. Der er også brug for, at børn og unge vælger at uddanne sig i naturvidenskabelige og tekniske retninger og stiller sig til rådighed for arbejdsmarkedet. Denne problemstilling stod centralt, da den daværende regering udkom med deres nationale naturvidenskabsstrategi tilbage i 2018 (Regeringen, 2018). Særligt siden har grundskolens naturfag fået stor opmærksomhed.

Allerede i forbindelse med folkeskolereformen i 2013 blev et arbejde med udvikling af folkeskolens afgangsprøver igangsat. Moderniseringen involverede blandt andet indførelse af en fælles praktisk/mundtlig prøve på tværs af de tre naturfag fysik/kemi, biologi og geografi samt en selvrettende skriftlig udtræksprøve i fysik/kemi. Senere blev der også indført selvrettende skriftlige udtræksprøver i geografi og biologi.

Forud var gået en bevægelse mod en samtænkning af naturfagene med de nye fælles mål i 2009 (Rambøll, 2018) og et stigende fokus på udvikling af fire centrale naturfaglige kompetencer: at undersøge, at modellere, at kommunikere og at perspektivere. Kompetencebegrebet slog for alvor igennem i Danmark i forbindelse med indførelsen af de nye Forenklede Fælles Mål, hvor målene for naturfagene blev formuleret i kompetencetermer (Dolin, 2014). Den udvikling prægede moderniseringen af de naturfaglige prøver, og siden har der været et ønske om, at særligt de skriftlige udtræksprøver i endnu højere grad skal afprøve elevernes kompetencer.

Et konkret initiativ i den nationale naturvidenskabsstrategi involverer udvikling af de tre udtræksprøver på forsøgsbasis. Der blev lagt op til, at forsøget blandt andet kunne involvere udvidelse af prøvernes varighed og opgaveantal med det formål at afprøve eleverne bedre i både bredden og dybden. På den baggrund initierede Børne- og Undervisningsministeriet et 2-årigt forsøg med prøverne, som oprindeligt skulle være startet op i 2021. Både i sommeren 2021 og sommeren 2022 blev prøverne dog aflyst grundet covid-19-pandemien, hvorfor implementeringen af forsøgsprøverne først skete i forbindelse med afgangsprøverne i sommeren 2023.

VIVE har fået til opgave at evaluere forsøgsprøverne på opdrag af Styrelsen for Undervisning og Kvalitet (herefter STUK), som gennemfører forsøget. Resultaterne af evalueringen præsenteres i denne rapport.

Forsøgsprøverne

De skriftlige udtrækprøver, der afprøves i forsøget, omtales forsøgsprøver. De væsentligste nye elementer, der indgår i forsøgsprøverne, er 1) anvendelse af simuleringer og 2) mulighed for at angive korte tekstsvar.

Forsøgsprøverne består af to dele – del A og del B. Del A svarer til den eksisterende prøve, der består af selvrettende multiple choice-spørgsmål, dog med færre opgaver (trefjerdedele af den eksisterende prøve). Del B består af opgaver, hvor eleverne skal formulere korte, åbne tekstsvar. Ved disse opgaver kan der indgå naturfaglige simuleringer, video og animationer. Simuleringer er et interaktivt element i prøven som fx en undersøgelse fra fysik/kemi, hvor eleven kan titrere væsker og aflæse resultater og dermed forholde sig til egen empiri.

Sammenlagt har forsøgsprøverne en tidsramme på 90 minutter, hvor de nuværende udtrækprøver har en tidsramme på 60 minutter. Eleverne styrer selv, hvor lang tid de bruger på hver del af forsøgsprøverne, men de er designet, så det forventede tidsforbrug er 45 min. på hver del. Del A af forsøgsprøverne gennemført i sommeren 2023 bestod af mellem 54 og 58 delopgaver, mens del B bestod af 17 eller 18 delopgaver (herefter omtalt "items"). Alle elever får de samme opgaver i samme rækkefølge, og hvert år formuleres der nye opgaver af en opgavekommission for hvert fag.

Forsøgsprøverne afprøver ligesom de nuværende udtrækprøver eleverne i kompetencemålene og de faglige områder inden for færdigheds- og vidensområderne for hhv. biologi, geografi og fysik/kemi, som de er beskrevet i Fælles Mål. Både de eksisterende prøver og forsøgsprøverne gennemføres i udgangspunktet digitalt ved en computer, medmindre prøven aflægges på særlige vilkår.

Da forsøgsprøverne indeholder opgaver, der skal besvares med elevernes selvstændige formuleringer, bliver prøvens del B bedømt af et statsligt beskikket censorkorps. Eleverne får en samlet karakter for del A og del B.

125 folkeskoler og frie grundskoler deltog i forsøget i skoleåret 2022/2023. På 83 skoler blev mindst én klasse udtrukket til en af prøverne.

(Børne- og Undervisningsministeriet, 2023).

1.1 Formål

Det primære formål med evalueringen er at etablere et beslutningsgrundlag for, om forsøgsprøverne skal gøres permanente efter forsøgsperioden. Overordnet skal det således undersøges, om forsøgsprøverne lever op til ambitio-

nen om i højere grad end de eksisterende prøver at afprøve elevernes naturfaglige kompetencer. Evalueringen skal samtidig bidrage med viden, der kan bruges til yderligere at optimere forsøget og prøverne.

Formålet indfris gennem fire delanalyser:

1. **Forsøgsprøvernes kvalitet.** Denne delanalyse har til formål at undersøge forsøgsprøvernes kvalitet gennem et særligt fokus på prøvernes indhold. Analysen vil blandt andet svare på, om prøverne er konstrueret hensigtsmæssigt, om der er sammenhæng i sværhedsgrad på tværs af prøverne, og om forsøgsprøverne favoriserer særlige elevgrupper.
2. **Oplevelsen af forsøgsprøver.** I denne delanalyse undersøger vi, hvordan lærere og elever oplever forsøgsprøverne med særligt fokus på de nye elementer, skriftlige svar og simuleringer.
3. **Ændringernes betydning for undervisningen.** Denne delanalyse har til formål at undersøge, hvad prøveændringerne betyder for undervisningen og eleverne.
4. **Forsøget og oplevede barrierer.** Den afsluttende delanalyse fokuserer på lærernes oplevelse af at tage del i forsøget og ikke mindst de barrierer for forandring, de oplever.

1.2 Metode

Evalueringen baserer sig på et ganske omfattende datamateriale bestående af både prøvedata, registerdata, interviewdata og survey-data.

Prøvedata stammer fra forsøgsprøverne i fysik/kemi, biologi og geografi afholdt i forbindelse med afgangsprøverne i sommeren 2023. Vi har suppleret disse med data fra afgangsprøverne i dansk (læsning), dansk (retskrivning) og matematik (uden hjælpemidler). Vi benytter data fra disse prøver til at undersøge sammenhængen mellem dygtigheden i andre fag og resultaterne i forsøgsprøverne.

Registerdata om elevens forældres uddannelsesniveau og oprindelse samt elevens biologiske køn og standpunktskarakterer i 8. klasse er også anvendt. Disse data dækker også elever, der ikke er en del af forsøget, så det er muligt at lave relevante sammenligninger.

Interviewdata involverer i alt 12 fokusgruppeinterviews med naturfagslærere på 9. årgang samt 12 fokusgruppeinterviews med elever på samme årgang. Interviewene er gennemført ad to omgange på i alt 12 skoler i forbindelse med

eller i forlængelse af eksempelprøverne gennemført i henholdsvis oktober 2021 og december 2022. Herudover har vi data fra i alt ni individuelle lærerinterviews – tre lærere fra hvert naturfag – gennemført i løbet af efteråret 2023. Interviewene er planlagt og gennemført på en måde, der sikrer balanceret repræsentation af de tre fag og en god geografisk fordeling af skoler. Der indgår desuden både folkeskoler samt privat- og friskoler.

Survey-data består af data fra en spørgeskemaundersøgelse blandt naturfagslærere fra skoler, der deltager i forsøget, og som underviste i mindst ét naturfag i skoleåret 2022/2023 og således stod for den prøveforberedende undervisning. I alt 132 lærere gennemførte hele spørgeskemaet. På grund af den anvendte distributionsmetode, som gik gennem skolernes kontaktpersoner, kan der ikke beregnes en præcis svarprocent. Men vi ved, at der i 90 % af de deltagende skolers tilfælde er svar fra mindst én lærer.

Datagrundlaget uddybes i rapportens Del 2, hvor de anvendte analysemetoder også beskrives. Analysemetoderne introduceres også løbende gennem rapportens Del 1, hvor det er relevant.

1.3 Læsevejledning

Rapportens Del 1 er struktureret ud fra de fire delanalyser nævnt i indledningen. I Kapitel 2 præsenterer vi således resultaterne af analyserne af forsøgsprøvernes kvalitet. Disse suppleres i Kapitel 3 af beskrivelser af, hvordan lærere og elever oplever forsøgsprøverne, mens prøveændringernes betydning for undervisningen, eleverne og udtræksprøvernes sammenhæng med den fælles naturfagsprøve beskrives i Kapitel 4. I Kapitel 5 præsenterer vi, hvordan lærerne oplever at tage del i forsøget med fokus på oplevede barrierer.

I rapportens Del 2 beskriver vi data og metoder. Del 2 indeholder også et længere kapitel med supplerende analyseresultater relateret til Kapitel 2 om forsøgsprøvernes kvalitet.

2 Forsøgsprøvernes kvalitet

For at der kan træffes beslutning om prøvernes fremtid, er det helt centralt at blive klogere på, hvordan forsøgsprøverne fungerer, og om de har den fornødne kvalitet – altså om de lever op til deres formål og samtidig har de egenskaber, man vil forvente af prøver af deres slags. Det har vi undersøgt, og i dette kapitel præsenterer vi resultaterne.

For at kunne forstå resultaterne er det nødvendigt kort at beskrive logikken bag de gennemførte analyser, og hvilke ønskede egenskaber ved prøverne de hjælper os med at vurdere. Analyserne er opdelt i følgende fire temaer:

1. Prøvernes sværhedsgrad

En prøve kan kun anvendes, hvis den har en passende sværhedsgrad til at afdække forskelle i elevernes dygtighed. Vi undersøger derfor, hvordan prøverne passer til elevernes færdigheder, og hvordan prøvernes sværhedsgrad i den nye prøvedel, del B, adskiller sig fra sværhedsgraden i det eksisterende format, del A.

Undersøgelsesspørgsmål:

- Afprøver forsøgsprøverne elevernes naturfaglige kompetencer, sådan som det er målet med prøverne?
- Hvordan klarer eleverne sig i del A sammenlignet med del B?
- Er der sammenhæng i sværhedsgrad på tværs af prøverne i de tre fag?

Forsøgets repræsentativitet

Håbet med forsøget er at blive klogere på, hvordan forsøgsprøverne fungerer på tværs af danske 9. klasse-elever. Det er kun muligt, hvis de elever, der deltager i forsøget, faktisk ligner og dermed repræsenterer resten af landets 9. klasse-elever. Dette er af afgørende betydning, idet det betyder, at analysernes gyldighed ikke begrænser sig til specifikke elever, men derimod dækker hele gruppen af elever. Samtidig muliggør det undersøgelser af forskelle på tværs af elevgrupper.

Resultaterne viser overordnet set, at de knap 3.000 elever, som har gennemført en forsøgsprøve, er repræsentative for danske 9. klasse-elever i forhold til elevernes socioøkonomiske baggrund og faglige niveau. geografisk dækker de omkring 83 skoler, hvor mindst én klasse blev udtrukket til en forsøgsprøve i sommeren 2023, kommuner i hele landet. Læs mere om repræsentativitet i afsnit 8.1.

Forsøgsprøverne består af to dele: del A og del B. Nogle elever har kun gennemført én af delene. Vi finder dog ingen statistisk signifikante forskelle i elevbaggrund mellem de elever, der har gennemført del A, og de elever, der har gennemført del B.

2. Prøvernes grundlæggende prøveegenskaber

Udover at prøvernes sværhedsgrad skal passe til elevernes dygtighed, er det vigtigt, at prøverne har en række grundlæggende egenskaber. Fx skal prøvernes delelementer gerne hænge sammen på en måde, så de kan bruges til at måle elevernes dygtighed i faget og kun det.

Undersøgelsesspørgsmål:

- Er prøverne konstrueret hensigtsmæssigt i forhold til opgavetyper, svarmuligheder, tidsforbrug mv.?
- Afprøver forsøgsprøverne elevernes naturfaglige kompetencer, sådan som det er målet med prøverne?
- Har brug af simuleringer og korte tekstsvar betydning for, om elevernes naturfaglige kompetencer afprøves?

3. Forskelle på tværs af elevgrupper

Det er forventeligt, at forskellige elevgrupper oplever prøver forskelligt af forskellige årsager. For eksempel vil elever, der klarer sig bedre i andre fag, typisk også klare sig bedre i naturfagene. Det afgørende er dog, om disse forskelligheder er større i den nye del, del B, især da denne prøveform har en anden form, der involverer tekstsvar, som kan opleves forskelligt på tværs af elever.

Undersøgelsesspørgsmål:

- Favoriserer forsøgsprøverne bestemte elevgrupper?
- Skyldes eventuelle forskelle i, hvordan elevgrupper klarer sig, forsøgsprøvernes konstruktion?

4. Bedømmelse af prøvesvar

Da forsøgsprøverne kræver en censorbedømmelse, er det afgørende, at de har et format, hvor de giver grundlag for ensartede bedømmelser på tværs af bedømmere.

Undersøgelsesspørgsmål:

- Er censorerne tilstrækkeligt enige i deres bedømmelser af prøvebesvarelser?

Analyseresultaterne præsenteres nedenfor. På grund af analysernes kompleksitet samler vi løbende op på hovedresultaterne i bokse i starten af hvert af dette kapitels underafsnit. Herudover beskriver vi kort kapitlets datagrundlag

og de anvendte metoder i boksen nedenfor. Uddybninger heraf kan findes i rapportens Del 2.

Boks 2.1 Data og metode

Datakilder

- De statistiske analyser bygger primært på data fra STUK på elevernes resultater i forsøgsprøverne i 2023. Disse data indeholder følgende informationer:
  - Antal opnåede point for hvert item i både del A og del B i de tre fag biologi, fysik/kemi og naturfag for hver elev.
  - For del B indeholder data det allokerede pointantal for hvert item særskilt for hver af de to censorer.
- For at kunne sammenholde resultaterne i forsøgsprøverne med elevernes færdigheder i andre fag har vi også anvendt data på det samlede antal point i afgangsprøverne i dansk læsning, dansk retskrivning og matematik (uden hjælpemidler) fra STUK.
- Data fra STUK er blevet leveret til Danmarks Statistik. Danmarks Statistik har via uni-login genereret pseudoanonymiserede cpr-numre, som vi har brugt til at sammenkoble data fra forsøgsprøverne med registre fra Danmarks Statistik, som giver information om forældrenes oprindelse, forældrenes uddannelsesniveau, elevens biologiske køn og elevens standpunktskarakterer i 8. klasse. Data fra Danmarks statistik indeholder desuden denne information for elever, der ikke deltog i forsøgsprøverne, og muliggør dermed, at vi kan undersøge om, eleverne i forsøgsprøverne er repræsentative for hele elevårgangen.

Metoder og anvendt software

- Vi har anvendt to type analyser i dette kapitel. For det første har vi gennemført statistiske analyser af gennemsnit og fordelinger. Desuden har vi undersøgt betingede og ubetingede sammenhænge, primært ved brug af mindste kvadraters metode-estimeringer af lineære modeller. For det andet har vi gennemført seks Rasch-analyser. Vi har gennemført Joint Maximum Likelihood-estimering af den simple dikotome Rasch-model for del A i de tre fag. For del B har vi estimeret Rasch-modeller af typen Partial Credit Model (PCM).
- De statistiske analyser og den primære databehandling er gennemført med programmet Stata version 18.0. Rasch-analyserne er primært gennemført ved brug af pakken TAM i programmet R version 4.2.0.

2.1 Prøvernes sværhedsgrad

Som nævnt ovenfor er det essentielt, at prøver har en passende sværhedsgrad for effektivt at kunne vurdere elevernes dygtighed. Prøver, som er for lette og alle elever scorer maksimumpoint på, giver ikke indsigt i elevernes individuelle dygtighed eller svagheder. Tilsvarende kan man heller ikke skelne mellem dygtige og mindre dygtige elever, hvis en prøve er for svær, og ingen elever opnår point. En velbalanceret prøve kendetegnes i stedet ved, at elevernes gennemsnitlige score ligger midt på spektret af mulige point, samtidig med at prøven differentierer mellem elever i enderne af spektret. Ideelt bør få elever score enten nul eller maksimumpoint.

Ved forsøgsprøverne er det endvidere vigtigt at sammenholde sværhedsgraderne på tværs af den nye del, del B, og den eksisterende del A for på den ene side at belyse, om prøverne supplerer hinanden i deres sværhedsgrad, og på den anden side belyse sammenhængen mellem, hvordan eleverne klarer sig i de to dele. Hvis sammenhængen er perfekt, således at elevernes resultat i del A perfekt forudser deres resultat i del B, bidrager del B ikke med ny viden. Hvis der omvendt ikke er nogen sammenhæng, vil det tyde på, at de to dele måler dygtigheden på helt forskellige områder. Det er derfor vigtigt også at belyse, hvordan sammenhængen mellem de to prøver er.

Konkret har vi undersøgt følgende:

- Elevernes point i forhold til det maksimale antal opnåelige point
- Fordelingen af point blandt eleverne
- Sammenhængen i elevernes point mellem del A og del B.

Analyserne viser samlet set, at prøvernes sværhedsgrad passer til elevernes dygtighed, og at den nye del, del B, supplerer den eksisterende del, del A, godt ved at være bedre egnet til at identificere de meget dygtige elever.

Boks 2.2 Afsnittets undersøgelsesspørgsmål og hovedresultater

Er der sammenhæng i sværhedsgrad på tværs af prøverne i de tre fag?

- Eleverne har generelt nemmere ved biologiprøven i både del A og del B, som de både har den højeste gennemsnitlige score i, generelt har en højere score på de enkelte items i, og hvor de har svaret rigtigt på den højeste andel items.
- Der er mere spredning i elevernes resultater i fysik/kemi-prøven end i biologi- og geografiprøven.
- Eleverne har sværest ved fysik/kemi-prøven, når vi ser på de samme faktorer. Det gælder i både del A og del B.

Hvordan klarer eleverne sig i del A sammenlignet med del B?

- Eleverne har generelt sværere ved prøvernes del B, målt på hvor mange point de får i gennemsnit sammenlignet med det maksimale antal point.

Afprøver forsøgsprøverne elevernes naturfaglige kompetencer, sådan som det er målet med prøverne?

- Der er en stærk sammenhæng mellem elevernes resultater i del A og del B, hvilket indikerer, at den nye prøvedel, del B, afprøver de samme naturfaglige kompetencer.

2.1.1 Del A, biologiprøven og simuleringsopgaver er nemmest

Analyserne af opnåede point i forsøgsprøverne viser, at eleverne klarer sig bedst i biologiprøven, og at del A er nemmere end del B. I Tabel 2.1 præsenteres de gennemsnitlige point opnået i hvert fag og i hver delprøve. I del A opnår eleverne i gennemsnit mellem 66 % og 73 % af det maksimalt mulige antal point. I kontrast hertil ligger gennemsnittet for del B mellem 51 % og 61 % af det maksimale pointtal. Både i del A og del B fremstår biologi som det fag, hvor eleverne generelt opnår den største andel af mulige point, hvilket indikerer, at

denne prøve relativt set er den nemmeste af prøverne. Omvendt scorer eleverne lavest i fysik/kemi-prøven. Alle forskellene er statistisk signifikante på 5-%'s niveau.

Tabel 2.1 Pointgennemsnit i hver delprøve i hvert fag

	Del A			Del B		
	Gennemsnit	Standard-afvigelse	Procentvis andel point ud af mulige point	Gennemsnit	Standard-afvigelse	Procentvis andel point ud af mulige point
Biologi	39,4	5,5	73 %	22,7	6,1	61 %
Fysik/ kemi	36,8	7,4	66 %	21,4	8,9	51 %
Geografi	40,3	7,1	69 %	21,4	5,7	58 %

Anm.: Gennemsnittet i del B er fundet på baggrund af censor 1's vurdering. Forskellene er også statistisk signifikante.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Zoomer vi ind på de enkelte items i prøverne, ser eleverne ud til at klare sig bedst i de items, hvor de har skullet anvende simuleringer. Hvilke specifikke items, der er tale om, kan ses i Figur 8.8. Et nærmere kig på disse delopgaver viser, at det specifikt er de simuleringsopgaver, hvor eleverne udelukkende skal følge de tilgængelige instruktioner i at anvende simuleringen og derefter aflæse de korrekte oplysninger, der er nemme. Omvendt viser samme Figur 8.8 nemlig, at eleverne har sværere ved de simuleringsdelopgaver, hvor eleverne ikke direkte skal anvende simuleringen.

2.1.2 Eleverne fordeler sig på hele pointskalaen

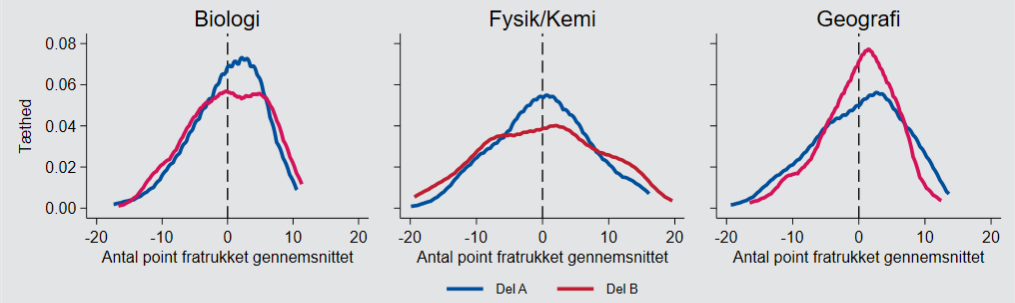
Ser vi nærmere på fordelingerne af det totale antal point i Figur 2.1, kan vi se, hvad vi også kan aflæse i Tabel 2.1, at spredningen i biologi og fysik/kemi er større i del B end i del A. Dette kan vi se, fordi fordelingerne for del B (de røde linjer i de tre figurer) er fladere og bredere end for del A. Dette er dog omvendt i geografi. Figuren illustrerer også, at eleverne, som nævnt ovenfor, klarer sig bedst i biologiprøven. Det indikeres af, at pointfordelingen for del A (den blå linje) i biologi er mere højreskæv end fordelingerne i del A i de andre to fag.

Figur 2.1 viser også, at pointfordelingerne i høj grad følger en normalfordeling. Det vil sige, at der ligger flest elever omkring midten af fordelingen og færre i kanterne. Prøverne fanger således elever på alle niveauer. Fordelingerne for del B er dog en smule fladere med flere elever ude i enderne af fordelingerne i

biologi og især i fysik/kemi. Dette indikerer, at denne del af prøven er bedre i stand til at identificere de meget dygtige og de mindre dygtige elever, end del A er. I geografi ser vi det modsatte mønster, hvor eleverne i højere grad klumper sig sammen i midten i del B end i del A. Sammenligner vi fordelingerne på tværs af fagene, ser det ud til, at det omvendte mønster i geografi delvis skyldes, at pointfordelingen i den eksisterende del, del A, er noget fladere i geografi end i de to andre fag. Med andre ord ligner fordelingen i del A i geografi i højere grad fordelingen for del B i de to øvrige fag.

Figur 2.1 Pointfordelingen i del A og del B

Figuren viser pointfordelingen for del A og del B i biologi, fysik/kemi og geografi. Pointfordelingerne for de to dele er centreret ved at fratrække gennemsnittet for nemmere sammenligning.



Anm.: X-aksen viser antal point fratrukket den gennemsnitlige score i hver del. De elever som her ligger på 0 point i del A har således fået 39,4 point og i del B 22,7 point (jf. Tabel 2.1). Fordelingen af point kan også ses som histogrammer i Figur 8.2, Figur 8.3 og Figur 8.4.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

2.1.3 God sammenhæng i sværhedsgrad på tværs af del A og B

Vi har også undersøgt sammenhængen mellem, hvordan eleverne klarer sig i de to dele. Det har vi gjort af to grunde. For det første for at undersøge, om de nye prøver, del B, afprøver de relevante naturfaglige kompetencer. Vi vil forvente en stærk sammenhæng mellem de eksisterende prøver, del A, og de nye prøver. Hvis der ikke er en sådan sammenhæng, tyder det på, at prøverne måler forskellige kompetencer. For det andet undersøger vi sammenhængen på tværs af prøvedelene for at belyse, om prøverne komplementerer hinanden. Hvis resultaterne i prøvedelene er perfekt korreleret, bidrager den nye prøvedel ikke med ny viden. Hvis der derimod er lidt variation på tværs af prøverne, således at en elev, der klarer sig godt i del A, kan forventes at klare sig mindre godt eller meget godt i del B, vil det tyde på, at prøverne supplerer hinanden. En stærk, men ikke perfekt sammenhæng mellem, hvordan eleverne

klarer sig i del A og B, vil derfor både være tegn på, at prøverne måler de korrekte kompetencer, og at de supplerer hinanden. Analyserne viser, at sammenhængen er stærk, men ikke perfekt.

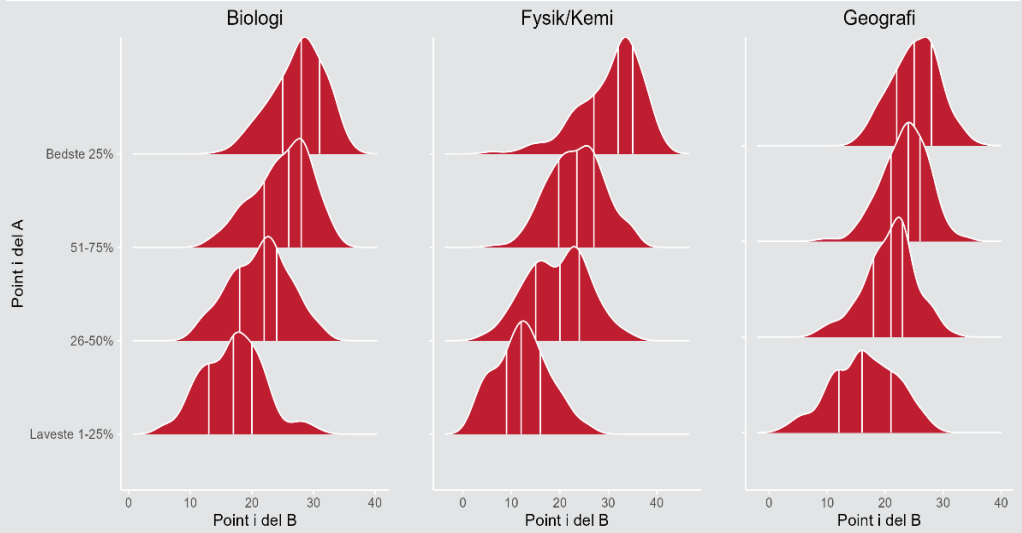
Figur 2.2 viser fordelingen af antal point i del B for fire grupper af elever defineret ved, hvordan de klarede sig i del A. Overordnet kan vi se, at fordelingen af antal point i del B ligger længst til højre for de bedste 25 % af elever i del A i alle fag. Det vil sige, at de elever, der klarer sig bedst i del A, også klarer sig bedst i del B. Vi kan dog også se en del overlap på tværs af grupperne. Det indikerer en vis mobilitet, forstået som at en elev kan bevæge sig fra at tilhøre den svageste (eller stærkeste) gruppe i del A til at tilhøre den bedste (eller svageste) gruppe i del B. Denne mobilitet tyder på, at den nye del, del B, giver ny viden og tillader nogle elever at vise deres dygtighed i et andet prøveformat, muligvis fordi den anden del ikke passede til dem, eller fordi de på grund af tilfældigheder ikke opnåede et resultat i del A, som svarer til deres dygtighed. Samtidig er der en stærk sammenhæng mellem resultaterne i del A og resultaterne i del B, som tyder på, at delene afprøver samme underliggende dygtighed.

Konkret er mobiliteten størst i geografi og mindst i fysik/kemi. Således lå ingen af de elever, der var i gruppen af de 25 % højest scorende elever i fysik/kemiprøvens del A, blandt de 25 % højest scorende elever i del B (fordelingen nederst i midten). Tilsvarende lå blot 3 % af dem, der var blandt de 25 % højest scorende elever i del B, blandt de 25 % lavest scorende i del A (fordelingen øverst i midten).

Blandt de 25 % elever, der lå i toppen i del A i biologi (fordelingen øverst længst til venstre), lå 56 % også i toppen i del B, og kun 2 % lå i de laveste 25 % i del B. Omvendt blandt de 25 %, der klarede sig dårligst i del A i biologi (fordelingen nederst længst til venstre), lå blot 3 % blandt de bedste 25 % højest scorende i del B.

I geografi ser vi, at 5 % af dem, der var blandt de lavest scorende 25 % i del A, er blandt de 25 % højest scorende i del B (fordelingen nederst til højre), og 4 % af de højest scorende 25 % i del A lå blandt de 25 % lavest scorende i del B (fordelingen øverst til højre).

Figur 2.2 Pointfordelingen i del B for kvartiler af pointfordelingen i del A



Anm.: De horisontale streger angiver kvartilerne, således at 25 % af observationerne ligger mellem to streger.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

2.2 Grundlæggende prøveegenskaber

For test og prøver er der, som nævnt indledningsvis, en række grundlæggende egenskaber og antagelser, som gerne skal være opfyldt. Hvis alle disse prøveegenskaber er opfyldt, er det muligt at drage nogle meget generelle konklusioner om prøvernes validitet og robusthed. Det vil typisk være lettere at opfylde disse grundlæggende antagelser for prøver, der undersøger meget præcist definerede kompetencer på en ensartet måde. I udviklingen af prøver vil der derfor ofte være en afvejning mellem at opnå disse egenskaber og samtidig afprøve forholdsvis brede kompetencebegreber (som for eksempel et helt naturfag) på forskellige måder (som for eksempel ved simuleringer og korte tekstsvar). Selvom man ønsker at prioritere den sidste del, er det naturligvis stadig vigtigt at undersøge de grundlæggende prøveegenskaber for at vurdere, om omkostningen ved at teste brede kompetencebegreber og anvende forskellige testmetoder er for høj. I dette afsnit præsenterer vi derfor resultaterne af vores analyse af fem grundlæggende egenskaber.

For at give læserne et samlet overblik og klæde dem på til at læse underafsnittene beskriver vi nedenfor, hvilke egenskaber vi har undersøgt, samt hvorfor og

hvordan. I underafsnittene udfoldes de væsentligste begreber og logikker yderligere.

1. Prøvernes interne konsistens

Prøverne består af en række spørgsmål eller items. For at prøverne samlet set giver et meningsfuldt svar på elevernes dygtighed inden for det pågældende fag, er det vigtigt, at disse items pålideligt måler det samme underliggende koncept – i dette tilfælde elevernes naturfaglige kompetencer. Et ekstremt scenarie er, hvor alle items måler noget helt forskelligt, og hvor et rigtigt svar på ét item er helt uafhængigt af et rigtigt svar på et andet item, fordi de måler dygtigheden på vidt forskellige områder. Dette er naturligvis ikke ønskeligt, idet prøven dermed ikke ville give en samlet, meningsfuld vurdering af elevens dygtighed inden for det pågældende fag. Hvis items er internt konsistente, tyder det derimod på, at den samlede score udgør en meningsfuld vurdering af elevens dygtighed inden for faget.

Vi har målt intern konsistens med Cronbachs Alpha.

2. Lokal uafhængighed

Intern konsistens, som beskrevet ovenfor, medfører, at et rigtig svar på ét spørgsmål er korreleret med et rigtigt svar på et andet spørgsmål. Hvis en elev svarer rigtigt på ét spørgsmål, er det et positivt tegn på vores vurdering af elevens dygtighed, og derfor vil vi også forvente, at eleven er mere tilbøjelig til at svare rigtigt på et andet spørgsmål inden for samme fag. Det er dog vigtigt, at denne sammenhæng kun skyldes elevens dygtighed, og hvis vi tager den dygtighed ud af billedet, bør der helst ikke være en sammenhæng længere. Det kaldes lokal uafhængighed, fordi svaret på ét spørgsmål skal være uafhængigt af svaret på et andet spørgsmål, når vi har taget højde for elevernes dygtighed. Vi kan for eksempel forestille os, at et spørgsmål beder eleverne svare på, hvad a+b+c giver, og at et andet spørgsmål beder eleverne svare på, hvad a+b giver. I det tilfælde vil et rigtigt svar på det sidste spørgsmål være en forudsætning for svaret på det første spørgsmål, og spørgsmålene er ikke længere uafhængige og bidrager ikke med uafhængig ny viden om elevens dygtighed.

For at tage højde for elevernes dygtighed har vi anvendt en Rasch-model og undersøger den lokale uafhængighed ved at betragte de parvise korrelationer af residualerne fra Rasch-modellen. I Rasch-modellen vurderes elevernes sandsynlighed for at svare rigtigt på et spørgsmål alene ud fra spørgsmålets sværhedsgrad og elevens dygtighed. Den variation, der er tilovers, når de to faktorer er taget i betragtning, kaldes residualerne. Vi kan derfor bruge residualerne til at undersøge, om der er korrelationer på tværs af spørgsmål, som ikke skyldes sværhedsgraden eller elevernes dygtighed.

3. Endimensionalitet

Den interne konsistens dækker over, at items bevæger sig i samme retning. Det vil sige, at dygtige elever inden for et fag svarer rigtigt på de fleste spørgsmål, og de mindre dygtige svarer rigtigt på færre spørgsmål. Som beskrevet ovenfor vil et ekstremt scenarie være, hvis hvert item måler noget vidt forskelligt. I en sådan situation vil prøven måle forskellige dimensioner. Er item 1 fx et spørgsmål om kemi, mens item 2 er et spørgsmål om historie, vil man sige, at prøverne måler to forskellige dimensioner. Udover at belyse prøvernes interne konsistens vil vi derfor også direkte undersøge, hvor mange dimensioner hver prøvedel dækker over.

I undersøgelsen af prøvernes endimensionalitet har vi fulgt samme fremgangsmåde som i undersøgelserne af den lokale uafhængighed og tager først højde for elevernes færdigheder ved brug af en Rasch-model. Vi har så undersøgt, hvor mange dimensioner der er tilbage i residualerne ved hjælp af en principalkomponentanalyse.

En principalkomponentanalyse anvendes normalt til at reducere antallet af dimensioner i data. Hvis man for eksempel har 20 variable, som man mener, måler færre underliggende egenskaber eller dimensioner, kan man bruge en principalkomponentanalyse til at belyse, hvor mange dimensioner der er tilbage i data. Givet at vi ved hjælp af Rasch-modellen har taget højde for elevernes dygtighed og spørgsmålenes sværhedsgrad, forventer vi, at residualerne kun dækker over "rent støj", og derfor bør der ideelt set ikke være meningsfulde dimensioner tilbage.

4. Spørgsmålenes stabilitet

Den fjerde egenskab, vi undersøger, er hvert items stabilitet. Et godt prøvespørgsmål er stabilt i den forstand, at hvis spørgsmålet er svært, bør svage elever sjældent svare rigtigt på det. Hvis et item ikke er stabilt, kan vi ikke forudsige, om en elev svarer rigtigt på det ud fra elevens dygtighed.

Vi undersøger også items stabilitet ved brug af Rasch-modellen, idet vi ser på variationen i elevernes besvarelser på hvert item ud fra elevernes dygtighed og spørgsmålets estimerede sværhedsgrad. Dette kaldes infit og outfit.

5. Ensartet vurdering på tværs af elevgrupper

En femte vigtig egenskab ved en god prøve er, at den virker ensartet på forskellige elevgrupper. Hvis det ikke er tilfældet, siges prøven at have Differential Item Functioning (DIF). Særligt grundet introduktionen af skriftlige svar, som er en betydelig ændring, der medfører en specifik bekymring for, om sær-

lige elevgrupper vil blive stillet dårligere, har vi dog valgt at præsentere resultaterne relateret til denne egenskab i sit eget afsnit. Det giver os bedre mulighed for at udfolde resultaterne tilstrækkeligt.

Rasch-modellen

Vi har valgt at fokusere på disse fem centrale egenskaber, da de er en forudsætning for, at prøverne er i tråd med Rasch-modellen, som sikrer, at prøvernes resultater er meningsfulde og robuste. Hvis dette ikke var tilfældet, eksempelvis fordi antagelserne om konsistens og endimensionalitet ikke var opfyldt, ville man kunne have to elever med samme resultatet i en prøve i et givent fag, selvom de i virkeligheden har vidt forskellige dygtigheder. De har kun opnået den samme score, fordi prøven afprøver vidt forskellige færdigheder. Den ene elev opnåede primært sit resultat på baggrund af elevens færdigheder på ét af de områder, mens den anden primært opnåede sit resultat på baggrund af dennes færdigheder på et andet område.

Det er dog vigtigt at bemærke, at vi her bruger Rasch-modellen og ovenstående fem egenskaber som analyseredskab i den forstand, at vi har valgt ikke at modificere prøverne eller Rasch-modellen for at opnå en bedre overensstemmelse mellem prøverne og de fem egenskaber. En sådan justering kunne for eksempel være at fjerne nogle items fra prøven og genestimere Rasch-modellen eller ved at udvide Rasch-modellen til at tage højde for en afvigelse (for eksempel DIF). Det har vi gjort af tre grunde.

For det første fordi prøvernes specifikke struktur og spørgsmål vil variere fra år til år. En specifik justering er derfor ikke meningsfuld på sigt. For det andet mener vi, at det i stedet er mere meningsfuldt for det fremtidige arbejde med prøvernes form at have information om de specifikke områder, hvor prøverne afviger, og om disse afvigelser er større i den nye del end i den eksisterende del. For det tredje vil en sådan justering i høj grad reducere analysernes gennemsigtighed, fordi vi for hvert fag og hver del vil have justeret både items og model.

En ulempe ved at primært at bruge Rasch-modellen som analyseredskab er, at modellens egenskaber ikke vil være fuldstændig opfyldt, givet at modellen ikke fitter perfekt. For eksempel vil estimater af elevdygtighed ikke nødvendigvis altid være i tråd med Rasch-modellens krav, og analysen af residualerne vil derfor afvige fra den gængse Rasch-analyse. Her vil vi dog fremhæve, at en Rasch-model, i den simpleste form, blot er en sandsynlighedsmodel med såkaldte fixed effects (items) og random effects (elever), hvor vi udfører en analyse af residualerne.

Boks 2.3 Afsnittets undersøgelsesspørgsmål og hovedresultater

Er prøverne konstrueret hensigtsmæssigt i forhold til opgavetyper, svarmuligheder, tidsforbrug mv.?

- **Intern konsistens:** For 2 ud af de 6 prøver tyder Cronbachs Alpha på acceptabel intern konsistens, og for de resterende fire prøver tyder den på god intern konsistens. Opgaverne er derfor konstrueret på en måde, så de lever op til de almindelige krav til konsistens.
- **Lokal uafhængighed:** I del B i fysik/kemi og geografi er andelen af items med tegn på brud på antagelsen om, at spørgsmålene er uafhængige af hinanden, værd at undersøge nærmere. For de andre delprøver er andelen af items med afvigelser så lav, at det sandsynligvis kan skyldes tilfældigheder.
- **Stabilitet**: Stabilitet måler spørgsmålenes forudsigelighed. På tværs af mål for stabilitet og fag er der ikke noget entydigt mønster, der peger på, at udfordringerne er størst i den nye del, del B.
- **Samlet:** Prøverne udviser fin intern konsistens, men med hensyn til lokal uafhængighed og stabilitet er der specifikke items, der kunne undersøges nærmere. Der er ikke noget, der tyder på, at den nye del generelt har dårligere prøveegenskaber end den eksisterende del.

Afprøver forsøgsprøverne elevernes naturfaglige kompetencer, sådan som det er målet med prøverne?

- **Endimensionalitet:** For 4 ud af 6 prøver er der tegn på brud med Raschmodellens antagelse om, at delprøverne kun måler en underliggende dimension af kompetencer, men disse udfordringer er ikke større i del B end i del A. Generelt kunne der være tegn på, at flere af prøverne afdækker flere naturfaglige kompetencer, men de kan godt stadig være inden for samme fag. For eksempel kan det tænkes, at fysik og kemi udgør to dimensioner, men de færdigheder afprøves i en prøve.

Har brug af simuleringer og korte tekstsvar betydning for, om elevernes naturfaglige kompetencer afprøves?

- Der er generelt ikke noget mønster i resultaterne, der peger på, at den nye prøvedel med korte tekstsvar og simuleringer generelt har hverken dårligere eller bedre prøveegenskaber end den eksisterende del.

2.2.1 Intern konsistens: Alle delprøver har mindst en acceptabel grad af intern konsistens

Vi har først undersøgt, om prøverne er internt konsistente i den forstand, at prøverne måler samme underliggende dimension af dygtighed. Skal en prøve eksempelvis måle dygtighed i biologi, skal alle items afprøve biologifærdigheder. Vi ser derfor på, om de rå spørgsmålsbesvarelser går i samme "retning". Med det mener vi, at rigtige spørgsmål på en delmængde af items også er korreleret med de resterende items. Det, at spørgsmål bevæger sig i samme retning (i.e. alle bevæger sig mod rigtige besvarelser, eller alle bevæger sig mod forkerte besvarelser), kaldes intern konsistens.

Konkret bruger vi målet Cronbachs Alpha til at måle denne interne konsistens. Cronbachs Alpha giver en værdi mellem 0 og 1, og højere værdier indikerer større intern konsistens. Normalvis anses værdier over 0,7 for acceptable og værdier over 0,8 for gode (Taber, 2018). For vigtige test og prøver, der har praktisk betydning for deltagerne, kan man dog også kræve en lidt højere grad af konsistens. Resultaterne er vist i Tabel 2.2.

Vi kan se, at for alle dele og begge fag ligger Cronbachs Alpha tydeligt over den acceptable grænse, og i 4 ud af 6 tilfælde er værdien i det gode område.

Tabel 2.2 Cronbachs Alpha

		Alpha	95-%'s konfidensinterval	
			Nedre	Øvre
Biologi	A	0,73	0,70	0,75
	B	0,80	0,78	0,82
Fysik	A	0,82	0,80	0,84
	B	0,86	0,85	0,87
Geografi	A	0,81	0,79	0,82
	B	0,75	0,73	0,78

Anm.: Tabellen viser Cronbachs Alpha og 95-%'s konfidensintervallerne.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

De to delprøver, hvor Cronbachs Alpha kun er acceptabel, er i del A i biologi og i del B i geografi. Der er derfor ingen tegn på, at der er større udfordringer med intern konsistens i den nye del, del B. Samlet set tyder det derfor på, at der ikke er udfordringer med mangel på intern konsistens.

2.2.2 Lokal uafhængighed: Tegn på udfordringer, der er værd at undersøge nærmere

Det er vigtigt, at sammenhængen i besvarelserne på tværs af prøverne kun bliver drevet af elevernes dygtighed og ikke af en konstrueret sammenhæng på tværs af items, som ikke opfanges af dygtighed. I indledningen af dette kapitel illustrerede vi denne sammenhæng ved, at et item spurgte til summen af a+b+c, mens et andet spørgsmål spurgte om resultatet af a+b. Det er forventeligt, at besvarelserne på de to items er korreleret, fordi de afprøver samme færdighed (addition), og de er dermed konsistente. Men denne korrelation skal kun komme fra elevernes dygtighed til addition og ikke fra, at det andet spørgsmål faktisk var et delspørgsmål af det første (a+b indgår begge steder). Derfor har vi undersøgt sammenhængen i besvarelserne på tværs af spørgsmål, hvor vi tager højde for elevernes dygtighed. I det konkrete eksempel svarer det til at tage højde for elevernes færdigheder i addition og dermed kun fange den sammenhæng, der skyldes, at spørgsmålene er konstrueret på en sådan måde, at de er korreleret ud over elevernes dygtighed.

For at tage højde for elevernes dygtighed og spørgsmålenes sværhedsgrad har vi estimeret en Rasch-model separat for hvert fag og for hver del.[1] På baggrund af de estimerede Rasch-modeller kan vi udtrække residualerne, som er den variation, der er tilbage i besvarelserne, når vi har taget højde for sværhedsgrader og elevernes dygtighed. Vi kan belyse den lokale afhængighed ved at beregne den parvise korrelation mellem alle kombinationer af items i hvert fag og hver del. For eksempel ser vi på residualkorrelationen i besvarelserne mellem item 9.1 og item 9.2 i del A i biologi. Korrelationerne er såkaldte Pearson-korrelationskoefficienter. De har en værdi mellem -1 og 1, hvor -1 angiver, at de er perfekt negativt korreleret, og 1, at de er perfekt positivt korreleret. Jo tættere værdierne er på nul, desto svagere er korrelationen.

Tabel 2.3 viser alle korrelationer, der er større end 0,2 (i absolutte enheder), som er den gængse nedre grænse for relevante korrelationer (Christensen et al., 2017). Generelt kan vi af tabellen se, at der er relativt få af de parvise korrelationer, der er større end 0,2 i absolutte enheder. Givet at vi her undersøger mange parvise korrelationer, er det ikke overraskende, at enkelte af dem er større end 0,2. I biologi del A er det 0,56 % af alle parvise sammenligninger og i del B er det 0,65 %. I fysik/kemi er det henholdsvis 0,71 og 4,58 %, og i geografi er det henholdsvis 0,67 og 2,21 %. De relativt høje andele i del B i fysik/kemi og i geografi kan med fordel undersøges nærmere.

Som det fremgår af Tabel 2.3, er flere af korrelationerne positive og indenfor samme hierarkiske struktur. For eksempel er der en forholdsvis stærk positiv korrelation i besvarelserne til item 9.4 og 9.2 i del A i biologi og mellem item

[1] Se afsnit 8.3.1 for flere detaljer om Rasch-analysen.

20.3 og 20.2 En løsning på dette brud er at estimere Rasch-modellen særskilt for del 9 og del 20 (og for de øvrige overordnede dele). Denne sammenhæng kunne opstå, fordi vi Rasch-modellen antager en underliggende dimension af dygtighed, men hvis de forskellige dele afprøver forskellige underdimensioner, så kan der opstå parvise korrelationer, som vi ser det her.

Hvor de positive korrelationer inden for underdelene er forventelige, er positive korrelationer på tværs af underdele og negative korrelationer helt generelt mere overraskende og foruroligende. For eksempel ser vi en negativ korrelation mellem besvarelserne på item 20.4 og item 20.1 i del A i biologi og en negativ korrelation mellem item 1.6 og 2.4 i del B i biologi. Givet at vi undersøger over 1.000 parvise kombinationer i del A og over 100 parvise kombinationer i del B, er det ikke overraskende, at der er flere korrelationer, der er negative og over 0.2 på tværs af den hierarkiske prøvestruktur. Det kan blot skyldes tilfældigheder. Men det er værd at undersøge de specifikke items, der er oplistet i Tabel 2.3, for at afklare om korrelationerne kan skyldes spørgsmålenes konstruktion, og om de bør forbedres. Dette gælder især for del B i fysik/kemi og geografi, hvor andelen er noget højere end i de andre delprøver.

Tabel 2.3 Lokal afhængighed

Del A					Del B			
	Items		Korrelation	Andel (i %)	Items		Korrelation	Andel (i %)
Biologi	9.4	9.2	0,55	0,07	2.4	1.6	-0,22	0,65
	20.3	20.2	0,42	0,14				
	20.4	20.1	-0,33	0,21				
	5.2	5.1	0,31	0,28				
	9.3	9.1	-0,30	0,35				
	17.3	17.2	-0,24	0,42				
	15.3	15.1	0,24	0,49				
	18.3	18.2	-0,22	0,56				
Fysik/kemi	8.3	8.1	0,46	0,06	1.7	1.6	0,47	0,65
	10.2	10.1	0,44	0,13	3.5	3.4	0,37	1,31
	19.2	19.1	0,42	0,19	2.2	1.5	-0,27	1,96
	8.3	8.2	0,41	0,26	2.2	1.6	-0,25	2,61
	8.2	8.1	0,40	0,32	1.7	1.5	0,23	3,27
	19.3	19.1	0,36	0,39	3.4	3.3	0,23	3,92
	21.3	21.2	0,29	0,45	2.2	1.7	-0,22	4,58
	21.2	19.1	0,29	0,52				

Del A					Del B			
	Items		Korrelation	Andel (i %)	Items		Korrelation	Andel (i %)
	16.2	16.1	-0,23	0,58				
	19.3	19.2	0,22	0,65				
	5.2	5.1	-0,21	0,71				
Geografi	20.4	20.1	0,36	0,06	3.6	3.5	0,34	0,74
	18.2	18.1	-0,34	0,12	3.7	1.2	-0,27	1,47
	20.4	20.2	0,32	0,18	3.3	1.5	-0,21	2,21
	20.2	20.1	0,28	0,24				
	14.3	14.2	0,26	0,30				
	2.4	2.2	0,23	0,36				
	10.3	10.1	-0,23	0,42				
	4.4	4.1	0,22	0,48				
	14.4	14.3	0,21	0,54				
	4.3	4.1	0,21	0,60				
	20.4	20.3	0,20	0,67				

Anm.: Tabellen viser alle parvise Pearson-korrelationskoefficienter på residualerne fra Rasch-modellen. Andelene angiver den kumulative andel af samtlige parvise kombinationer. Bemærk, at del A består af flere items, og selvom der er flere items med korrelationer over cutoff i den del, er det relativt mindre sammenlignet med del B, som det fremgår af kolonnen 'Andel (i %)'.

Tabellen viser alle korrelationer, der er større end 0,2 og signifikante på et 5 %'s niveau med en Bonferroni-korrektion.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

2.2.3 Endimensionalitet: Der er tegn på, at prøverne ikke opfylder antagelsen om endimensionalitet, men problemet er ikke større i den nye del af forsøgsprøverne

I det foregående afsnit præsenterede vi resultaterne af analysen af, om besvarelserne af de enkelte items i prøverne er korreleret, ud over hvad der kan forklares ved elevernes dygtighed og spørgsmålenes sværhedsgrad. Det gjorde vi ved at se på residualerne, som netop er den variation i besvarelserne, der er tilovers, når vi har taget højde for dygtighed og sværhedsgrad. Vi har også undersøgt en anden væsentlig egenskab ved residualerne i form af antallet af dimensioner, som de repræsenterer.

Når vi betragter et antal variable vil det ofte være tilfældet, at de repræsenterer et mindre antal underliggende uobserverbare dimensioner. For eksempel kan fire variable, der dækker over henholdsvis husstandsindkomst, forældrenes uddannelsesniveau, husets størrelse og antal biler i husstanden, repræsenterer et mindre antal underliggende dimensioner, for eksempel kulturel og

finansiel kapital. For at nå til en sådan konklusion kan man anvende en såkaldt principalkomponentanalyse, hvor man undersøger, hvor mange "komponenter" ens variable repræsenterer. I dette tilfælde vil vi undersøge antallet af komponenter i residualerne. Hvis principalkomponentanalysen peger på, at residualerne repræsenterer en eller flere meningsfulde komponenter, tyder det på, at variationen, der er tilbage i residualerne, ikke kun skyldes støj og tilfældigheder, men at der er flere dimensioner tilovers i besvarelserne, hvilket vil være et brud med antagelsen om endimensionalitet.

Det er en grundlæggende egenskab ved en Rasch-model, at en prøve kun måler én underliggende dimension. Problemet ved, at en prøve dækker flere dimensioner, bliver blandt andet tydelig i aggregeringen af elevernes resultater. Prøverne anvendes til at give en samlet vurdering af elevernes færdigheder, og hvis man i den proces har aggregeret scorer sammen, som repræsenterer flere underliggende dimensioner, er det for eksempel uklart, om en middel samlet score repræsenterer, at eleven er middelgod på to underliggende dimensioner eller god på den ene og mindre god på den anden. I den konkrete situation vil der sandsynligvis kunne opstå brud på antagelsen om endimensionalitet, idet prøverne afdækker relativt brede kompetencer (som for eksempel fysik/kemi). En mulighed er derfor at teste flere underliggende Rasch-modeller for hvert delspørgsmål. Men da vi her primært anvender Rasch-modellen som analyseredskab, har vi i stedet estimeret en Rasch-model for hver prøvedel og afrapporterer resultaterne med fokus på, om udfordringerne er større i den nye prøvedel.

I Tabel 2.4 viser vi de fem største egenværdier for de to dele i hvert fag. Egenværdierne svarer til komponenter eller dimensioner, og egenværdiens størrelse indikerer, hvor stor en del af variationen i data komponenten forklarer. Hvis komponenterne er store, tyder det derfor på, at de forklarer en stor del af variationen i elevernes besvarelser, efter vi har kontrolleret for elevernes dygtigheder og spørgsmålenes sværhedsgrad. Vi kan se, at i 4 ud af 6 tilfælde er den største egenværdi større end 2 og derfor et tegn på brud med antagelsen om endimensionalitet. Der er dog ikke tegn på, at problemet er større i del B end i del A – nærmere tværtimod.

En mulig fremgangsmåde er derfor at estimere en Rasch-model for hver underdel af prøverne, men da det primære formål med analyserne er at vurdere deres kvalitet og identificere mulige problemstillinger, vil vi i stedet fokusere på at identificere mulige kilder til bruddet med endimensionalitet. I Tabel 2.3 så vi, at flere af de stærke parvise sammenhænge er inden for samme spørgsmålsblok, hvilket indikerer en hierarkisk struktur. Det er derfor plausibelt, at flerdimensionalitet dækker over, at de enkelte blokke måler lidt forskellige kompetencer, hvilket ikke er overraskende givet et overordnet ønske om at afprøve forholdsvis brede kompetencebegreber.

Tabel 2.4 De største egenværdier fra en principalkomponentanalyse af residualerne

	Del	De fem største egenværdier				
		1	2	3	4	5
Biologi	A	2,12	1,79	1,72	1,64	1,57
	B	1,69	1,45	1,30	1,18	1,16
Fysik	A	2,49	2,26	1,90	1,68	1,54
	B	2,16	1,72	1,34	1,23	1,20
Geografi	A	2,16	1,95	1,71	1,62	1,57
	B	1,63	1,48	1,46	1,25	1,17

Anm.: Tabellen viser de fem største egenværdier for en PCA af residualerne fra Rasch-modellen.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

2.2.4 Stabilitet: Der er ikke tegn på, at spørgsmålene i den nye del er mere ustabile end spørgsmålene i den eksisterende del

Udover at prøvernes items ideelt set bør være konsistente, lokalt uafhængige og også kun måle en underliggende dimension, bør de også være stabile. Stabilitet refererer her til, at et spørgsmål er forudsigeligt. Hvis et spørgsmål er svært, vil vi forvente, at en større andel af de dygtige elever svarer rigtigt på spørgsmålet end blandt de mindre dygtige. Ved et ustabilt item har vi svært ved at forudsige, om en elev svarer rigtigt på spørgsmålet, selvom vi både har et mål for elevens dygtighed og et mål for spørgsmålets sværhedsgrad.

Vi har målt spørgsmålenes sværhedsgrad ved igen at bruge Rasch-modellen og ud fra den udregne et mål for, hvor "stabile" spørgsmålene er. Stabilitet dækker over, om besvarelserne er som forventet. Ved spørgsmål, som er estimeret til at være svære, forventer vi, at det primært er dygtige elever, der svarer rigtigt, og kun få af de mindre dygtige. Omvendt for spørgsmål, som er estimeret som lette, forventer vi, at de fleste svarer rigtigt. Hvis den faktiske besvarelse afviger meget fra vores forventninger, er det et tegn på, at spørgsmålene er ustabile.

Konkret har vi brugt to mål til at vurdere stabiliteten af spørgsmål. For begge mål gælder det, at vi har sammenholdt elevernes faktiske besvarelser med, hvad vi forventer ud fra elevens dygtighed. Så hvis et spørgsmål er estimeret til at have en sværhedsgrad på 0,5 logit, og elevens dygtighed er estimeret til at være 0,5, så vil vi forvente, at der er 50 %'s chance for, at eleven svarer rigtigt på spørgsmålet, fordi det netop passer præcis til elevens dygtighed. Hvis der så er langt flere (eller langt færre) af eleverne med en dygtighed på 0,5 logit, der svarer rigtigt på det spørgsmål, er der tegn på dårlig stabilitet, fordi

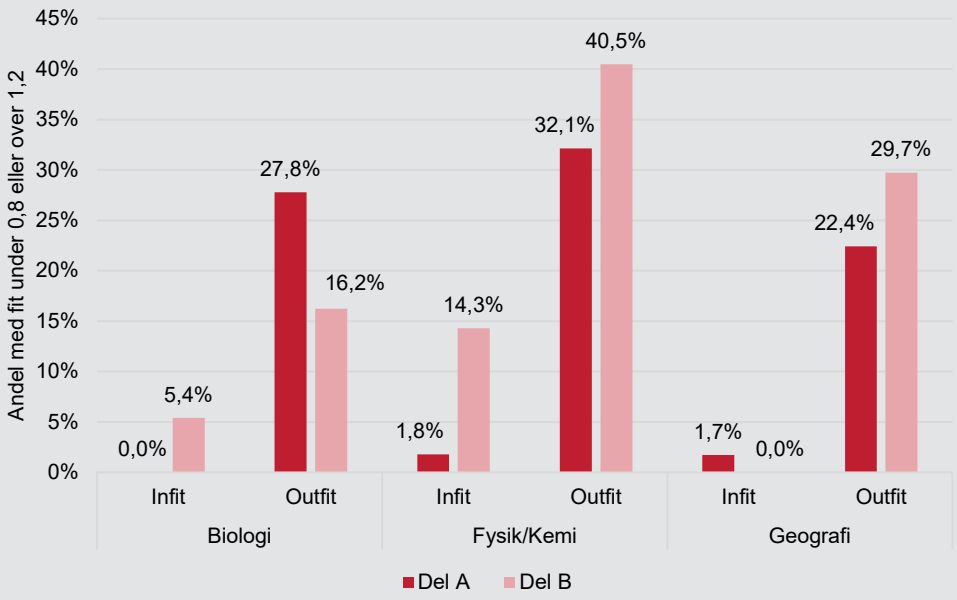
spørgsmålet ikke giver stabile besvarelser. De to mål, vi anvender, kaldes infit og outfit. Forskellen på de to mål ligger i, hvor meget vægt vi lægger på de enkelte elever. I outfit-målet har alle elever lige stor betydning, mens infit-målet udregner stabiliteten ved primært at betragte besvarelser fra elever, hvis estimerede dygtighed ligger tæt på spørgsmålets sværhedsgrad. I ovenstående eksempel vil en elev med en estimeret dygtighed på 0,5 logit altså få langt større betydning end en elev med en estimeret dygtighed på 1 logit.

I Figur 2.3 viser andelene af items med infit og outfit uden for det, der betegnes som normalområdet for high-stakes-prøver som disse (Wright & Linacre, 1994). For infit-målet er der generelt ret få items uden for normalområdet med undtagelse af del B i fysik/kemi. For outfit-målet er andelen langt højere på tværs af fagene og delene.

Mens det er værd at undersøge items med dårligt fit enkeltvis (se afsnit 8.3.2 for en oversigt over de specifikke items og deres afvigelser), fremgår det også af Figur 2.3, at der ikke er noget mønster, der tyder på, at den nye del, del B, er mere ustabil end del A. Betragter vi infit-målet, er andelen størst i del B i biologi og fysik/kemi, men ikke i geografi. Betragter vi outfit-målet, er andelen af items, der vurderes ustabile, højest i del B i biologi, men ikke i fysik/kemi og geografi.

Betragter vi ikke kun andelene, men også den faktiske afvigelse, som vist i afsnit 8.3.2, er der specielt i del B nogle større afvigelser, som er værd at se nærmere på.

Figur 2.3 Items med infit og outfit uden for normalområdet



Anm.: Figuren viser andelen af items, hvor henholdsvis infit eller outfit er uden for normalområdet, som er defineret ved 0,8 til 1,2.

Kilde: VIVEs analyser på baggrund af data fra STUK og Danmarks Statistik.

2.3 Forsøgsprøvernes betydning for særlige elevgrupper

Ovenfor præsenterede vi resultaterne af analyser af prøvernes sværhedsgrad og egenskaber for den samlede gruppe af elever. I dette afsnit fokuserer vi i stedet på, om prøven virker forskelligt på tværs af elevgrupper.

Når en prøve ændres, kan ændringerne potentielt føre til, at bestemte elevgrupper stilles dårligere end før relativt til andre grupper. Fx skal eleverne selv formulere skriftlige svar i forsøgsprøverne, mens de i de eksisterende prøver blot skal vælge mellem mulige svar. Særligt denne ændring har både hos opdragsgiver og blandt de interviewede lærere skabt grund til bekymring. For vil den betyde, at det i højere grad end i de eksisterende prøver er elevernes danskkundskaber, der vurderes, end det er elevernes dygtighed i det pågældende naturfag? Det har derfor været et centralt spørgsmål at undersøge. Ge-

nerelt tyder vores analyser dog på, at ændringerne *ikke* stiller bestemte elevgrupper dårligere end ved de eksisterende prøver, og at der således ikke er grund til bekymring. Hvordan vi er nået frem til dette, uddybes nedenfor.

Vi præsenterer først resultaterne af undersøgelsen af, i hvilken grad elevernes færdigheder i dansk har betydning for, hvordan eleverne klarer sig i forsøgsprøverne. Dernæst ser vi på, om der er forskelle på tværs af elevbaggrund med hensyn til elevens biologiske køn, oprindelse og forældrenes uddannelse.

Frafaldne elever uden betydning for resultater

I alt har 2.676 elever besvaret forsøgsprøvernes del A og 2.589 elever forsøgsprøvernes del B i biologi, geografi eller fysik/kemi. Der er således ikke lige mange elever, der har besvaret de to dele. Der er 424 elever, der har besvaret del A, men ikke del B, og 337 elever, der har besvaret del B, men ikke del A. 2.252 elever har besvaret begge prøver. 2.246 af disse kan vi koble registerdata på, og det er denne gruppe elever, alle analyserne af forsøgsprøverne er lavet på baggrund af.

Vi har undersøgt om den gruppe af elever, som kun har svaret på én af de to delprøver, er anderledes end de elever, som har besvaret begge. Blandt dem, som kun har svaret på én af delprøverne, er andelen med ikke-vestlig oprindelse højere, karaktergennemsnittet i 8. klasse knap et halvt karakterpoint mindre og forældrenes antal års skolegang et par måneder lavere. At vi udelader disse elever er dog ikke noget, der ændrer resultaterne og de konklusioner, vi kan drage på baggrund af dem.

Boks 2.4 Afsnittets undersøgelsesspørgsmål og hovedresultater

Favoriserer forsøgsprøverne bestemte elevgrupper?

- Vi undersøger, om der er systematiske forskelle i, hvordan eleverne klarer sig i forsøgsprøverne med hensyn til elevens biologiske køn, elevens oprindelse og forældrenes uddannelse. Overordnet set er der ingen tegn på, at elevbaggrunden har en større betydning for elevens resultat i den nye del, del B, sammenlignet med den eksisterende del.

Skyldes eventuelle forskelle i, hvordan elevgrupper klarer sig, forsøgsprøvernes konstruktion?

- Givet de korte tekstsvar i de nye forsøgsprøver undersøger vi også specifikt, om danskfærdigheder har en særlig stor betydning for den nye delprøve. Det er forventeligt, at elever, som klarer sig bedre i dansk, også i gennemsnit klarer sig bedre i naturfag. Men overordnet set peger analyserne på, at danskfærdigheder ikke har en større betydning i den nye del, del B, end i den eksisterende del.

2.3.1 Elevernes danskfærdigheder har ikke større betydning i den nye delprøve

I analyserne har vores fokus været at undersøge, om del B (den nye del) er væsentlig anderledes end del A (den eksisterende del) med hensyn til, hvordan elever med forskellige færdigheder i dansk klarer sig i prøverne. I enkelte analyser har vi desuden undersøgt forskelle med hensyn til matematikfærdigheder som en slags benchmark. Vi har både anvendt simple statistiske analyser og Raschanalyser til at belyse forskelle med hensyn til elevbaggrund.

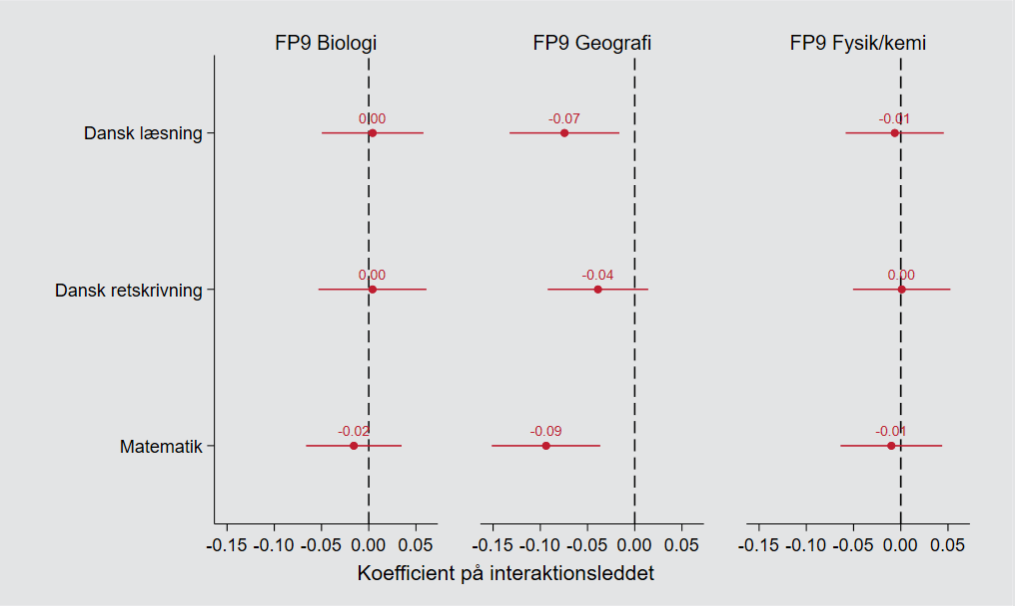
I Figur 2.4 viser vi resultaterne fra en statistisk analyse af, om færdigheder i dansk læsning, dansk retskrivning og matematik har en anden sammenhæng med del B sammenlignet med del A.[2] Vi kan se, at færdighederne i både dansk læsning og retskrivning ikke har en anden sammenhæng med resultaterne i

[2] Analysen er lavet ved at estimere en lineær model af det standardisererede resultat i hver af de to prøvedele på hver baggrundsparameter, en indikator for, om det er del A eller del B, og en interaktion mellem, om det er del B og elevens baggrund. Vi er her primært interesseret i interaktionsleddet. Hvis interaktionsleddet er stort statistisk forskelligt fra nul, tyder det på, at elevens baggrund har en systematisk anden betydning for del B sammenlignet med del A.

del B sammenlignet med del A. Det eneste sted, hvor vi ser en anden sammenhæng, er for matematik i geografi, hvor danskfærdighederne og matematikfærdighederne betyder *mindre* i del B sammenlignet med del A.[3]

Figur 2.4 Sammenhæng mellem elevernes resultater i øvrige prøver og den samlede score i prøvernes del A og del B

Figuren viser koefficienterne (de røde prikker) og 95-%'s konfidensintervallet (de røde linjer) for, om sammenhængen med del B er højere end sammenhængen med del A. En positiv sammenhæng betyder, at elevernes dansk/matematik-færdigheder har større betydning i del B. Rører konfidensintervallet de stiplede lodrette linjer, er sammenhængen *ikke* statistisk signifikant.



Anm.: For at gøre resultaterne sammenlignelige er resultaterne i prøverne blevet standardiseret til at have et gennemsnit på 0 og en standardafvigelse på 1 inden for hvert fag.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

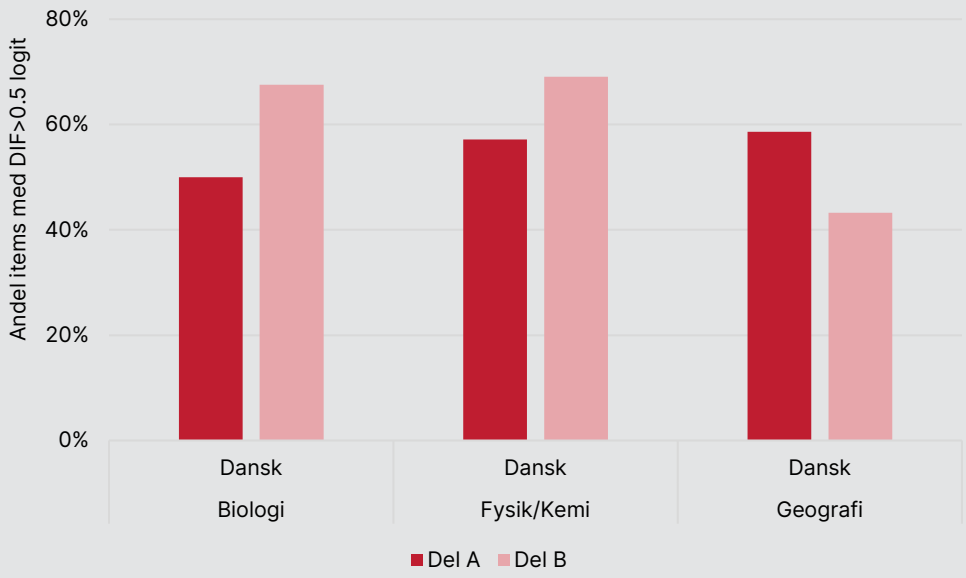
Hvor vi i ovenstående analyser har fokuseret på elevernes samlede score, kan vi ved hjælp af Rasch-analyserne også undersøge, om enkelte items opleves forskelligt for elever med dårligere eller bedre færdigheder i dansk. Sådanne forskelle betegnes som Differential Item Functioning (DIF) i Rasch-modellen. I Figur 2.5 viser vi andelen af items, hvor forskellen i den estimerede sværhedsgrad er større end 0,5 logit, som er et gængs cutoff for store forskelle på

[3] Hvor Figur 2.4 viser resultaterne for undersøgelsen, om færdigheder i andre fag har en anden betydning for gennemsnittet i del B sammenlignet med del A, viser Figur 8.21 tilsvarende analyse, hvor vi i stedet for gennemsnittet i andre fag ser på betydningen for at være blandt de bedste 20 %. Igen ser vi ikke noget tegn på, at sammenhængen er anderledes i del B end i del A.

tværs af grupper. Det er tydeligt, at der er en del items, som opleves forskelligt på tværs. Det er ikke overraskende, da elever, der opnår gode resultater i dansk, både er elever, der er særligt gode i dansk, og elever, der generelt får høje karakterer. Derfor er det ikke overraskende, at disse elever også vil klare sig bedre i naturfagene og derfor typisk vil opfatte de enkelte spørgsmål som "lettere". I biologi og fysik/kemi er andelen, hvor forskellen i den estimerede sværhedsgrad er på 0,5 logit eller mere, størst i del B, men i geografi er det omvendt. Der er derfor heller ikke her tegn på, at danskfærdighederne systematisk har en større betydning i den nye del.[4]

Figur 2.5 Andelen af items, hvor forskellen i de estimerede sværhedsgrader med hensyn til elevens danskfærdigheder er større end 0,5 logit

Figuren viser andelen af items, hvor forskellen i sværhedsgraden mellem grupperne er større end 0,5 logit-enheder for hver del og hvert fag.



Anm.: Grupperingen 'Dansk' opdeler børn i to grupper, afhængigt af om deres karaktergennemsnit i dansk læsning og dansk retskrivning er over 7.

Kilde: VIVEs egne beregninger baseret på data fra STIL og Danmarks Statistik.

[4] I Tabel 8.8 i afsnit 8.4 lister vi alle items, hvor forskellen i estimeret sværhedsgrad er større end 0,5 logit. Vi har valgt 0,5 som cutoff, da det er meget anvendt i litteraturen, men man kunne til sådanne high-stakes-prøver også anvende lavere cutoffs som 0,3 logit, men da vi primært sammenligner de to dele og pga. prøvernes udformning, hvor de ikke genbruges, har vi valgt at bruge et cutoff på 0,5.

2.3.2 Elevernes baggrund har heller ikke en større betydning i de nye delprøver

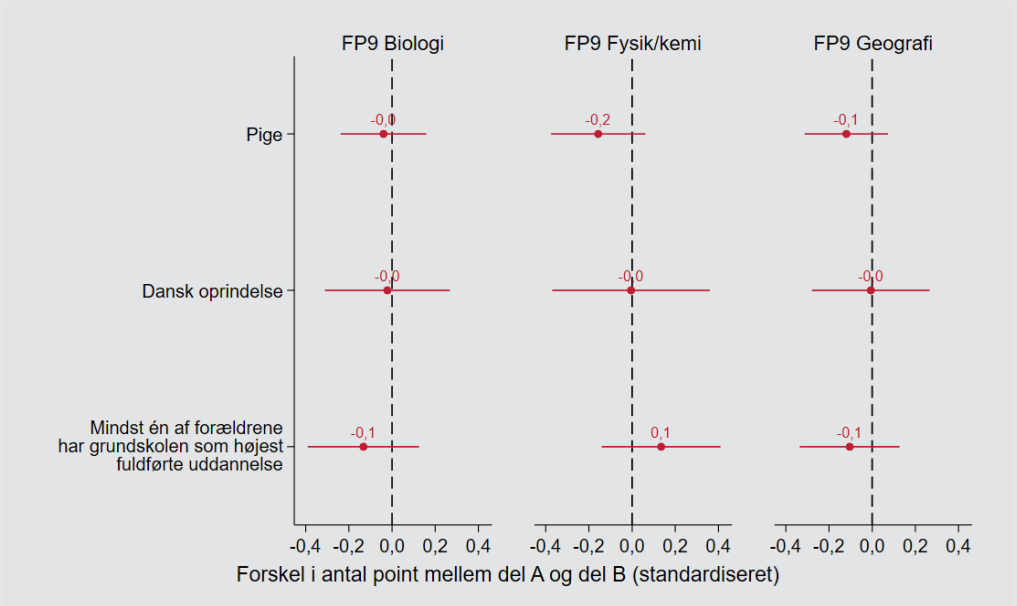
Udover at prøverne kan opleves forskelligt afhængigt af elevens færdigheder i dansk, kan prøverne også opleves systematisk forskelligt på tværs af forskellige elevbaggrunde. Det har vi undersøgt igen ved at fokusere på forskelle i den nye del, del B, sammenlignet med del A. Konkret har vi delt eleverne op på tre relevante parametre:

- Biologisk køn tildelt ved fødsel.
- Oprindelse: Eleven har dansk oprindelse, hvis begge forældre er født i Danmark.
- Forældres uddannelse: Grupperet ud fra, om mindst en af forældrene højst har gennemført grundskolen.

Resultaterne af statistiske analyser af, om elevens baggrund med hensyn til køn, oprindelse og forældrenes uddannelse har en større betydning i del B sammenlignet med del A, fremgår af Figur 2.6. Figuren illustrerer, at der ikke er tegn på, at elevernes baggrund har hverken større eller mindre betydning for, hvordan de klarer sig i del B sammenlignet med del A. Det kan vi se, fordi den røde, horisontale linje i alle tilfælde krydser den stiplede vertikale linje.

Figur 2.6 Sammenhæng mellem elevernes baggrund og den samlede score i prøvernes del A og del B

Figuren viser koefficienterne (de røde prikker) og 95-%'s konfidensintervallet (de røde linjer) for, om sammenhængen med del B er højere end sammenhængen med del A. En positiv sammenhæng betyder, at dansk/matematik-færdighederne har større betydning i del B. Rører konfidensintervallet de stiplede lodrette linjer, er sammenhængen *ikke* statistisk signifikant.



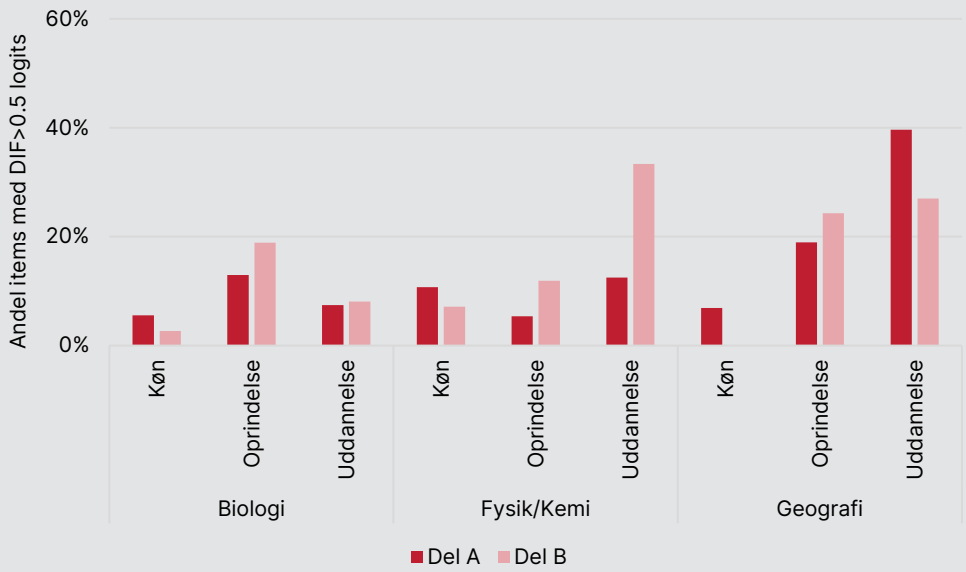
Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

I Figur 2.7 viser vi andelen af items i hver delprøve, hvor den estimerede sværhedsgrad er højere end 0,5 logit på tværs af de tre delgrupper. Generelt ser vi en langt lavere andel af items med store forskelle sammenlignet med Figur 2.5 men også her ser vi ikke noget systematisk tegn på, at forskellen er større i den nye delprøve.[5]

[5] I Tabel 8.8 i afsnit 8.4 lister vi alle items, hvor forskellen i estimeret sværhedsgrad er større end 0,5.

Figur 2.7 Andel items, hvor forskellen i estimeret sværhedsgrad er over 0,5 logit-enheder

Figuren viser andelen af items, hvor forskellen i sværhedsgraden mellem grupperne er større end 0,5 logit-enheder for hver del og hvert fag.



Anm.: Køn opdeler eleverne i grupper efter deres biologiske køn (dreng og pige). Oprindelse opdeler eleverne i to grupper, efter om begge deres forældre er født i Danmark. Uddannelse opdeler eleverne i to grupper, afhængigt af om mindst en forælder højst har gennemført grundskolen.

Kilde: VIVEs egne beregninger baseret på data fra STIL og Danmarks Statistik.

2.4 Pålidelighed på tværs af censorer

Da evalueringerne af forsøgsprøvernes del B udføres ved faglige bedømmelser gennemført af det beskikkede censorkorps, er det vigtigt at belyse, hvor stabile disse bedømmelser er. Det skal helst ikke være sådan, at en elevs resultat afhænger af, om eleven er heldig eller uheldig med at få en gavmild eller mindre gavmild censor. I stedet bør det være sådan, at to bedømmere af samme besvarelse typisk giver samme resultat.

Vi har undersøgt, om det er tilfældet i forsøgsprøverne. Det har været muligt, fordi hver besvarelse er blevet vurderet af to forskellige og tilfældigt udvalgte

censorer[6]. Resultaterne viser, at der generelt er stor enighed på tværs af censorerne. Resultaterne uddybes nedenfor.

Boks 2.5 Afsnittets undersøgelsesspørgsmål og hovedresultater

Er inter-rater-reliabiliteten tilstrækkeligt stor?

- Der er generelt stor enighed på tværs af censorerne. På items-niveau giver de to censorer samme pointantal i mellem 72 og 77 % af tilfældene på tværs af de tre fag.
- Betragter vi det samlede pointantal pr. elev, er bedømmernes uenighed mindre end 5 point for 85 % af eleverne i biologi. For fysik/kemi er det for 79 % af eleverne, og i geografi er det for 72 % af eleverne.
- Krippendorffs alpha er et mål for, hvor stor enigheden er, i forhold til hvad man vil forvente ved tilfældighed. Alpha-værdierne er her 0,86, 0,76 og 0,91 for henholdsvis biologi, fysik/kemi og geografi. Værdier på over 0,67 anses for at være acceptable.

2.4.1 Acceptabel enighed mellem censorer

I Tabel 2.5 viser vi for de tre fag fordelingen af overensstemmelse mellem to censorer på spørgsmålsniveau. Vi kan se, at i biologi er censorerne helt enige i antal point for 72 % af alle items. For 25 % af items er de uenige med ét point, og for 2 % af items er uenigheden større end 2 point. I fysik/kemi er der en lidt højere andel spørgsmål, hvor censorerne er helt enige, 77 %, men derimod er der også en lidt større andel, hvor uenigheden er større end 1 point, konkret 6,2 %. I geografi er censorerne helt enige om vurderingen af 73 % af alle items. På 6,5 % af spørgsmålene er uenigheden større end 1 point.

Det varierer dog fra item til item, hvor mange point der kan gives, og derfor har vi også undersøgt, hvor ofte uenigheden er så stor, som den kan være, givet det mulige antal point. Det er den i forholdsvis få tilfælde, nemlig i henholdsvis 1,7, 3,6 og 2,2 % af tilfældene, som Tabel 2.5 viser.

[6] I forbindelse med den endelige karaktergivning er én censorbedømmelse udvalgt tilfældigt for hver elev. Elevernes endelige bedømmelse, og dermed den opnåede karakter, er således tildelt på baggrund af én censors bedømmelse.

Tabel 2.5 Censorenighed på items-niveau

	100% enighed	Uenighed ml. censorer				
	0 point	1 point	2 point	3 point	4 point	Maksimal
FP9 Biologi	72,4 %	25,3 %	2,1 %	0,2 %		1,7 %
FP9 Fysik/kemi	77,4 %	17,6 %	3,9 %	0,8 %	1,5 %	3,6 %
FP9 Geografi	72,6 %	21,4 %	5,2 %	0,7 %	0,6 %	2,2 %

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Tabel 2.6 viser overensstemmelsen mellem censorer på den samlede score for en elev. I henholdsvis 14, 11 og 9 % af tilfældene er censorerne helt enige, og i langt de fleste tilfælde er enigheden inden for 4 point. For biologi er censorernes uenighed mindre end 5 point for 85 % af eleverne. I fysik/kemi og geografi er de tilsvarende andele 79 og 72 %.

Tabel 2.6 Censorenighed på samlede score

	Maks. antal point	100% enighed	Uenighed ml. censorer					
			1 point	2 point	3 point	4 point	5-10 point	>10 point
FP9 Biologi	37	13,9 %	22,7 %	21,0 %	17,6 %	9,9 %	14,4 %	0,6 %
FP9 Fysik/kemi	42	10,9 %	24,7 %	20,2 %	12,9 %	10,4 %	20,6 %	0,4 %
FP9 Geografi	37	9,3 %	19,9 %	17,6 %	14,3 %	10,6 %	27,5 %	0,8 %

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Det kan være vanskeligt at vurdere, om uoverensstemmelsen i den foregående tabel er stor eller lille. I Tabel 2.7 viser vi derfor Krippendorffs alpha, som er et mål for, hvor ofte censorerne er enige, i forhold til hvad vi vil forvente ved tilfældigheder. Hvis for eksempel to censorer er enige i hvert andet tilfælde, men at man givet det samtlige mulige antal bedømmelser kan beregne, at ved tilfældighed ville to censorer være enige i hvert andet tilfælde, så er pålideligheden ikke særlig høj, og Krippendorffs alpha vil være tæt på nul. Hvis det ved tilfældighed derimod er meget sjældent, at to censorer er enige, så vil en enighed på 50 % være meget høj, og Krippendorffs alpha være tæt på 1. Generelt anses en Krippendorffs alpha på over 0,8 for at være fremragende,

mens værdier over 0,67 er acceptable (Beckler et al., 2018). De beregnede Krippendorffs alpha fremgår af Tabel 2.7. Resultaterne indikerer, at enigheden mellem censorerne, der har vurderet fysik/kemi-prøven, er acceptabel, mens censorenigheden for de to andre prøver er fremragende.

Som det fremgår af boksen nedenfor, har censorerne da også været godt tilfredse med rettevejledningerne, som de har fulgt loyalt. Det kan bidrage til at forklare den gode pålidelighed i censorernes vurderinger.

Tabel 2.7 Krippendorffs alpha

	Alpha	95-%'s konfidensinterval	
Biologi	0,86	0,64	0,92
Fysik/kemi	0,76	0,39	0,87
Geografi	0,91	0,76	0,95

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Stor tilfredshed med rettevejledningerne

I forlængelse af afviklingen af eksempelprøverne i december 2022 gennemførte vi en spørgeskemaundersøgelsen blandt censorerne for at blive klogere på rettevejledningernes kvalitet, og hvordan de bliver brugt. 35 censorer svarende til 81 pct. besvarede spørgeskemaet. Resultaterne af undersøgelsen blev præsenteret for STUK på et efterfølgende møde.

Vi fandt overordnet set, at censorerne er meget tilfredse med rettevejledningerne og den understøttelse, der er tilgængelig for dem. 91 pct. er tilfredse eller meget tilfredse.

Censorerne mener, at rettevejledningerne har høj kvalitet, men at især overskueligheden og eksemplerne på elevbesvarelser kan forbedres. Et særligt udbredt ønske blandt censorerne er desuden, at de gives mulighed for at se opgaverne i samme format, som eleverne ser dem.

Videre udviser censorerne stor loyalitet over for rettevejledningerne, men der er enkelte eksempler på opgaver, hvor censorerne oplevede, at rettevejledningen ikke var tilstrækkelig til at give en bedømmelse. Det oplevede 29 pct. af censorerne. STUK har fået en liste over de konkrete opgaver.

Censorerne har været glade for den tilgængelige understøttelse, som bestod af et webinar et muligheden for at række ud til STUK omkring tvivlsspørgsmål. Få benyttede sig dog af sidstnævnte muligheder flere foretrækker et fysisk seminar frem for et webinar.

3 Oplevelse af forsøgsprøverne

Én ting er, hvad de statistiske analyser fortæller os om forsøgsprøverne, noget andet er, hvad lærere og elever kan fortælle os. Vi har derfor undersøgt, hvordan lærere og elever oplever forsøgsprøverne. Det har vi gjort via de tilgængelige interviewdata og spørgeskemaundersøgelsen blandt lærere. Resultaterne præsenterer vi i dette kapitel.

Konkret baserer kapitlet sig på følgende undersøgelsesspørgsmål:

- Hvad betyder brugen af simulering og skriftlige svar for oplevelsen af forsøgsprøverne?

Analyserne baserer sig på interviewdata og spørgeskemaundersøgelsen blandt lærere.

Lærerne hilser generelt ændringerne til prøverne velkomne. Både lærere og elever sætter pris på den større variation i opgavetyper. Særligt introduktionen af simuleringer fremhæves som noget positivt, mens skriftligheden opleves svær. Vi uddyber resultaterne nedenfor.

3.1 Et skridt i den rigtige retning

Lærerne er overvejende positive i deres vurderinger af forsøgsprøverne. Direkte adspurgt, hvor tilfredse de overordnet set er med forsøgsprøverne, svarer 65 %, at de er tilfredse eller meget tilfredse (se Figur 3.1). Til sammenligning giver 10 % af lærerne udtryk for utilfredshed, mens 25 % hverken er tilfredse eller utilfredse. En endnu større andel af lærerne – konkret 77 % – mener, at forsøgsprøverne overordnet set er et skridt i den rigtige retning. Hvorfor andelen, der overordnet er tilfredse med prøverne, ikke er lige så stor, kan vi ikke svare definitivt på. Men det kan hænge sammen med lærernes bekymringer omkring skriftlighedens betydning, som vi allerede tidligere har beskrevet. I hvert fald svarer 22 % af lærerne, at de er bekymrede for, hvordan prøveændringerne vil påvirke eleverne.

Alligevel er lærerne samtidig altovervejende positive i deres vurderinger af prøven. Det fremgår blandt andet af spørgeskemaresultaterne, der er illustreret i Figur 3.2. Eksempelvis vurderer et stort flertal på 80 % af lærerne, at forsøgsprøverne afprøver elevernes naturfaglige kompetencer – et vigtigt mål med ændringerne. Det har også været et tydeligt tema i interviewmaterialet.

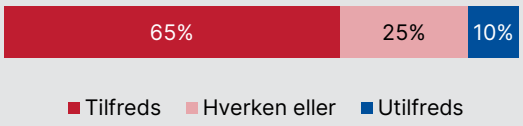
Lærerne er generelt af den holdning, at forsøgsprøverne er mere nutidige og bedre matcher undervisningen, og hvad eleverne forventes at kunne. Forsøgsprøverne opleves også mere virkelighedsnære, blandt andet fordi eleverne sjældent vil stå i virkelige situationer, hvor de blot skal forholde sig til verden gennem et multiple choice-spørgsmål, men snarere vil have brug for at forklare og argumentere.

> *Altså det er jo sjældent, at det bare er, at man kan sætte et kryds ved et eller andet svar, der er fortrykt.*
>
> **Lærer**

Lærerne oplever derfor også, at forsøgsprøverne i højere grad bidrager til elevernes dannelse, for som en lærer forklarer, "er det bare vigtigt, synes jeg, at man kan fortælle og forklare, hvad det er, der sker ude i den virkelige verden."

Ovenstående kommer ligeledes til udtryk i lærernes spørgeskemabesvarelser. Henholdsvis 70 og 71 % af lærerne mener, at forsøgsprøverne kræver, at eleverne udviser stor faglig selvstændighed, samt at forsøgsprøverne afprøver kritisk stillingtagen hos eleverne. 74 % mener, at forsøgsprøverne afprøver de gældende fælles mål.

Figur 3.1 Overordnet tilfredshed med forsøgsprøverne

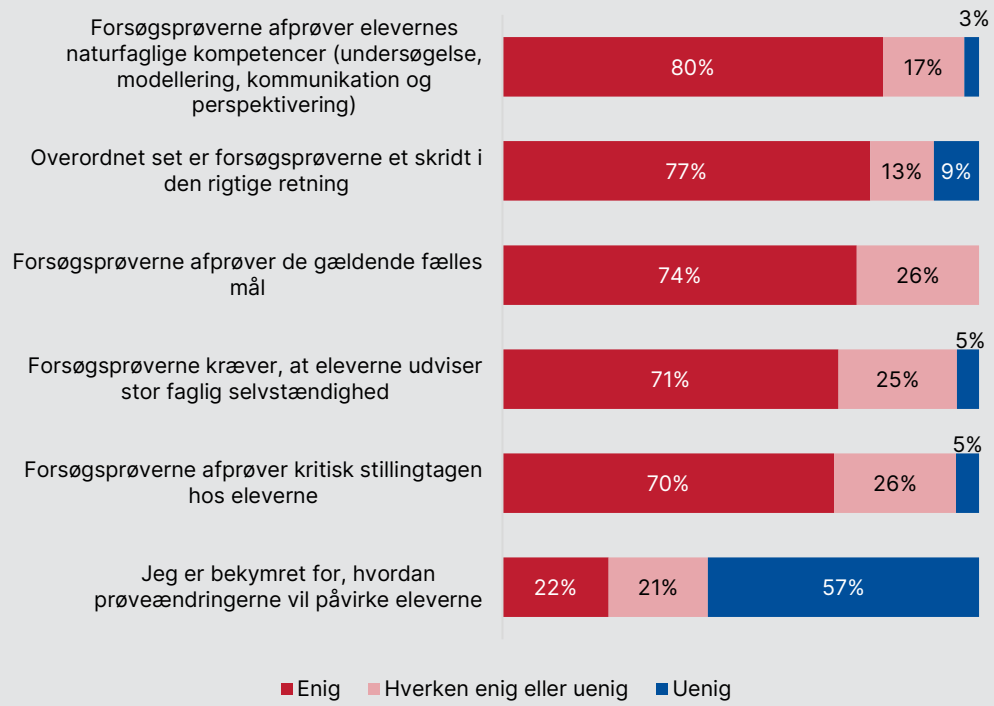


Anm.: n = 132.

Spørgsmålsformulering: 'Hvor tilfreds er du overordnet set med forsøgsprøverne?' Tilfreds = 'tilfreds' eller 'meget tilfreds'; utilfreds = 'utilfreds' eller 'meget utilfreds'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

Figur 3.2 Uddybet vurdering af forsøgsprøverne



Anm.: n = 121-130.

Spørgsmålsformulering: 'Hvor enig eller uenig er du i følgende udsagn?' Enig = 'enig' eller 'meget enig'; uenig = 'uenig' eller 'meget uenig'. 'Ved ikke'-svar er sorteret fra. Vedr. udsagnet 'forsøgsprøverne afprøver de gældende fælles mål' er svarkategorierne 'hverken enig eller uenig' og 'uenig' kollapset til 'hverken enig eller uenig' af hensyn til anonymitet.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

3.2 Skriftligheden opleves svær

Evalueringen har ikke involveret interview af elever efter afviklingen af forsøgsprøverne i sommeren 2023. Til gengæld gennemførte vi interviews med elever i forbindelse med eksempelprøverne i både skoleåret 2021/2022 og 2022/2023, som var opbygget på samme måde som forsøgsprøverne. Derudover gengav lærere under interviews i efteråret 2023, hvad deres elever havde fortalt dem om deres oplevelser i sommeren 2023. Disse gengivelser

beskriver ret entydigt, at eleverne har oplevet forsøgsprøverne som svære – særligt del B[7].

At del B opleves som svær, stemmer godt overens med, at vi i de statistiske analyser af sværhedsgrader præsenteret i afsnit 2.1 også finder, at del B er sværere end del A. Men interviewdata nuancerer resultatet med, at det er skriftligheden i prøven, der opleves svær.

> *Generelt oplevede eleverne prøven som svær, og en del af mine dygtige elever var meget nervøse for, hvordan det var gået. Men jeg opfatter det i hvert fald som et udtryk for, at de følte, de blev stillet nogle lidt sværere opgaver end det her med at krydse af. […] Nu skulle de jo rent faktisk formulere noget på skrift, og det er specielt de skriftlige svar, som nogle af dem oplever som svære.*
>
> **Lærer**

Eleverne selv fortalte om eksempelprøverne, at tilføjelsen af del B gør, at de ikke i lige så høj grad blot kan gætte sig til et godt resultat. De skal i højere grad bringe deres viden og kompetencer i spil og kende de relevante begreber, når de skal anvende dem skriftligt. Det gør, at prøven opleves sværere. Og det kan være svært for eleverne at gennemskue, hvornår et skriftligt svar er godt nok.

Lærerne giver udtryk for, at tilføjelsen af skriftlighed ikke kun gør, at prøverne opleves sværere – de bliver faktisk sværere. Helt som de statistiske analyser viser. Skriftligheden stiller større krav til eleverne, som i højere grad skal kunne argumentere og formulere sig med brug af korrekte faglige termer. Det betyder samtidig, at eventuelle huller i elevernes viden og kompetencer ikke lige så nemt skjules. Men det betyder også, at de dygtige elever bedre kan vise det via udfoldede svar.

Blandt lærerne er der dog bekymring for, at skriftligheden vil stille de sprogligt udfordrede elever dårligere. Dette både fordi dårlige danskkundskaber evt. kan stå i vejen for, at eleverne kan vise den dygtighed, de faktisk har, og fordi et dårligt formuleret svar fx kan være svært for censorerne at vurdere, selvom

[7] Ifølge lærerne er elevernes oplevelser af prøverne som svære dog ikke nødvendigvis ensbetydende med, at eleverne har klaret sig dårligt.

svaret måske er rigtigt. Som præsenteret under afsnit 2.3, indikerer vores analyser dog, at der ikke er stor grund til disse bekymringer.

3.3 Godt med variation og simuleringer

Både elever og lærere ser positivt på den større variation i opgavetyper – særligt simuleringsopgaverne er en god tilføjelse. Lærerne er glade for simuleringsopgaverne, fordi de oplever, at de giver bedre mulighed for at afprøve elevernes undersøgelseskompetence.

> *"De der simuleringer, de er fine. Det er ... det er godt tænkt, og ... Jamen, jeg synes, det er nogle fine undersøgelser, de sætter op, og jeg synes, det var fint ... men hele tanken om at spørge ind til nogle konkrete undersøgelser og lave nogle simuleringer synes jeg, fungerer godt.*
>
> **Lærer**

Simuleringsopgaverne giver også mulighed for, at eleverne ved at "tænke naturfagligt" kan ræsonnere sig frem til rigtige svar, selvom de ikke nødvendigvis står stærkt inden for det behandlede emne. Tilsvarende bliver der blandt eleverne givet udtryk for, at simuleringsopgaverne blandt andet er gode, fordi eleverne kan arbejde sig frem til et svar uden at skulle uddybe skriftligt. De opleves derfor også nemmere end de øvrige opgaver. I analyserne af prøverne kan vi da også se, at det netop er simuleringsopgaverne, som allerflest elever svarer rigtigt på (se afsnit 2.1.1).

Blandt eleverne er der derfor også en opfattelse af simuleringsopgaverne som en sjov, kærkommen pause fra resten af prøven. Det er også lærernes oplevelse, at eleverne er positive over for simuleringer. Flere lærere fremhæver således, at de oplever, at elevernes synes, at simuleringer er sjove og motiverende, når de arbejder med dem i undervisningen.

4 Ændringernes betydning

Ændres en prøve, vil det helt naturligt forventes også at lede til ændringer i undervisningen frem til prøven og dermed for eleverne og for sammenhængen til den fælles praktisk-mundtlige naturfagsprøve. Vi har derfor undersøgt, hvad tilføjelsen af skriftlige svar og simuleringer til prøverne betyder.

Konkret baserer kapitlet sig på følgende undersøgelsesspørgsmål:

- Hvad betyder brugen af simulering og skriftlige svar for undervisningen?
- Hvilke undervisningsformer og materialer anvender lærerne i undervisningen, når der skal undervises frem mod forsøgsprøverne?
- Hvordan tilrettelægges undervisningen, så der er sammenhæng i forberedelsen til de monofaglige skriftlige prøver og den fælles prøve?

Ligesom ovenfor baserer analyserne i kapitlet sig på interviewdata og spørgeskemaundersøgelsen blandt lærere. De samlede interviewdata er inddraget i analysen, men hovedfokus har været på lærerinterviewene gennemført i efteråret 2023, da disse havde et mere specifikt fokus på ændringernes betydning end de øvrige interviewdata.

Vi finder, at prøveændringerne fører til ret betydelige forandringer. Særligt simuleringer og skriftlighed fylder af naturlige årsager mere i undervisningen. Ændringerne betyder dog ikke det store for eleverne, og lærernes fokus er fortsat på den fælles naturfagsprøve, men sammenhængen mellem prøverne opleves bedre.

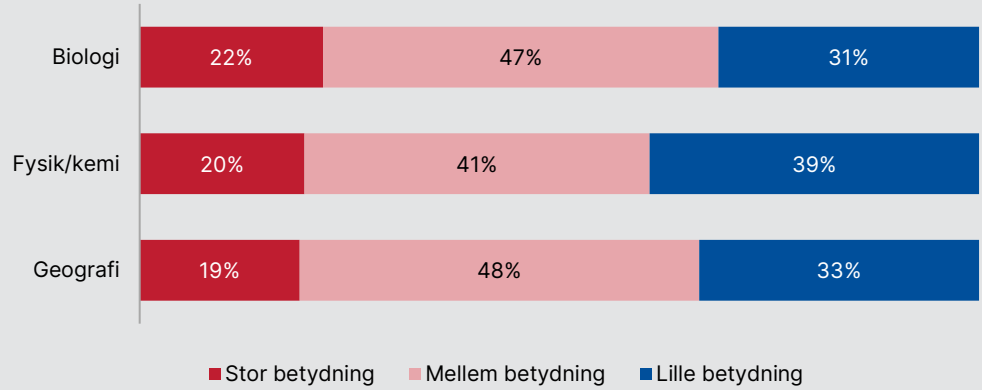
4.1 Ændringerne fører til forandringer

Når vi ser på tværs af data, tegner der sig et lidt mudret billede af prøveændringernes helt overordnede betydning for undervisningen. I interviewdata er der både lærere, der giver udtryk for, at de slet ikke har ændret noget i deres undervisning, og lærere, der giver udtryk for, at de har lavet ret betydelige ændringer. Selvom grundlaget er begrænset, synes der dog at være en tendens til, at tilpasningen af undervisningen så småt er ved at tage fart. I hvert fald gav de fleste lærere i forbindelse med de to første interviewrunder udtryk for, at de slet ikke eller kun i begrænset omfang havde ændret deres undervisningstilgang som følge af prøveændringerne. Men under den seneste interviewrunde – på den anden side af sommerens prøver – gav flere lærere udtryk for, at de ikke lavede nogen ændringer i skoleåret 2022/2023, men at de nu

kan se, at tilpasninger er nødvendige. Spørgeskemaundersøgelsen, som blev gennemført op til sommerferien 2023, tegner et tilsvarende billede af, at lærerne faktisk er i færd med at implementere ret betydelige ændringer i deres undervisning.

Som Figur 4.1 illustrerer, vurderer 19-22 % af lærerne, at forsøgsprøverne har haft stor betydning for deres undervisning, mens 41-48 % vurderer, at forsøgsprøverne har haft mellem betydning.

Figur 4.1 Forsøgsprøvernes samlede betydning for undervisningen



Anm.: n = 84-107.

Spørgsmålsformulering: 'I hvilken grad har forsøgsprøverne haft betydning for din undervisning i ...?' Stor betydning = 'i høj grad' eller 'i meget høj grad'; mellem betydning = 'hverken i høj eller lav grad'; lille betydning = 'i lav grad' eller 'i meget lav grad'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

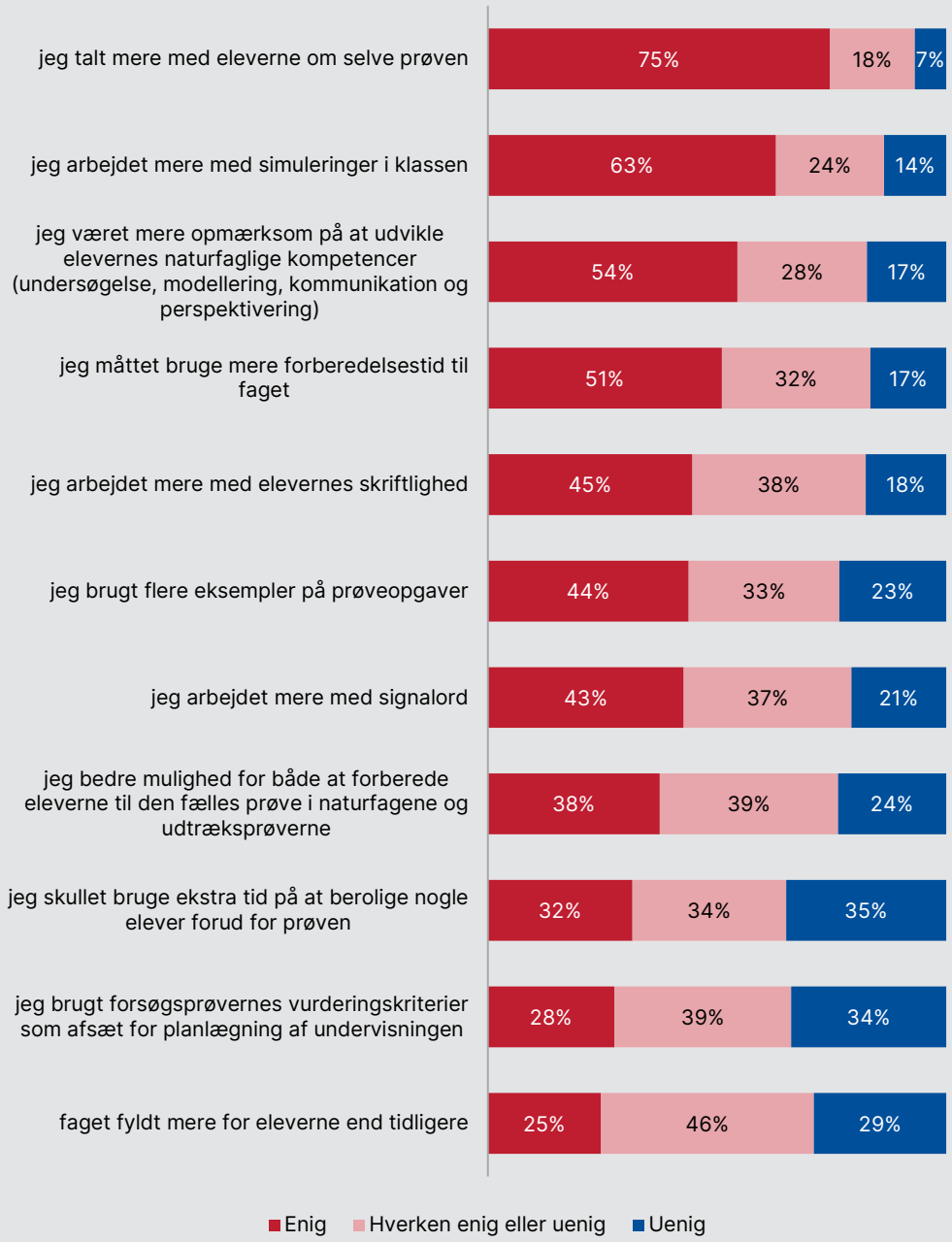
I spørgeskemaundersøgelsen har lærerne forholdt sig til en række konkrete udsagn omkring prøveændringernes betydning for deres undervisning. Både udsagn og resultater fremgår af Figur 4.2 nedenfor. Figuren fortæller blandt andet, at den største ændring er, at lærerne har talt mere med eleverne om selve prøven. Det er dog naturligt og forventeligt og derfor ikke så interessant i sig selv. Mere interessant er, at henholdsvis 63 og 45 % af lærerne er enige i, at de som følge af den nye prøveform har arbejdet mere med simuleringer og elevernes skriftlighed. I afsnittene nedenfor går vi mere i dybden med de resultater.

> *Altså jeg synes, at det vi har gjort, altså vi har øvet os i at skrive mere, og så har vi brugt tid på de her simuleringer. Det tror jeg faktisk er den væsentlige forskel i forhold til, hvad vi tidligere har gjort. Det er de to punkter*
>
> **Lærer**

Et andet interessant resultat er, at 54 % af lærerne svarer, at de som følge af prøveændringen har været mere opmærksomme på at udvikle elevernes naturfaglige kompetencer. Det stemmer meget godt overens med pointen i afsnit 3.1 om, at forsøgsprøverne har en bedre kobling til virkeligheden og fagenes mål end de eksisterende prøver. Fagenes mål relaterer sig netop til de fire naturfaglige kompetencer, og det virker derfor naturligt, at en bedre kobling mellem prøverne og fagenes mål vil lede til et større fokus på målene. Alligevel virker det overraskende, at over halvdelen af lærerne faktisk har øget deres fokus på kompetencerne allerede.

Et flertal af lærerne svarer også, at de har måttet bruge mere forberedelsestid til deres naturfag. De lærere, vi har interviewet, og som ikke har lavet ændringer i deres undervisning, fremhæver netop denne pointe – mangel på tid – som årsagen.

Figur 4.2 Uddybet betydning af forsøgsprøverne



Anm.: n = 142-148.

Spørgsmålsformulering: 'Hvor enig eller uenig er du i følgende udsagn? Som følge af den nye prøveform har ...' Enig = 'enig' eller 'meget enig'; uenig = 'uenig' eller 'meget uenig'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

Nedenfor gennemgår vi først, hvad skriftligheden har betydet for lærernes undervisning, og derefter simuleringernes betydning.

4.2 Skriftlighedens betydning for undervisningen

Flere lærere ser et behov for at arbejde med skriftligheden, men synes, det er svært at integrere i undervisningen. Lærerne peger især på, at der er behov for at øve argumentation, forklaringer og opbygning af svar, hvori der indgår fagbegreber. Forsøg og undersøgelser bliver fremhævet som oplagte til at øve skriftlighed i forbindelse med.

4.2.1 Behov for at arbejde med skriftlighed

Med tilføjelsen af del B ser flere lærere et behov for at arbejde med elevernes skriftlige svar i naturfagsundervisningen. Flere lærere påpeger dog også, at det halter med at integrere skriftligheden i undervisningen. Det er således en erkendelse hos flere, at skriftligheden har været for lidt i fokus i undervisningen frem mod forsøgsprøverne, og at det fremadrettet er nødvendigt at fokusere mere på.

> *Der er i hvert fald noget omkring den der sprogliggørelse, som vi lige skal tænke over, hvordan får vi det bygget ind i lidt højere grad end det, vi allerede gør*
>
> **Lærer**

En lærer beskriver det som, at det, de er gode til at øve mundtligt med eleverne, også skal have en skriftlig dimension: ”Vi kommer til at lave, har vi snakket om, nogle opgaver, hvor de skal prøve at forklare sig skriftligt. Vi har gjort det en del mundtligt, og det kan de egentlig godt, men hvor de også lige skal have det ned i nogle korte skriftlige formuleringer.”

Som nævnt ovenfor er 45 % af lærerne enige i, at de har arbejdet mere med elevernes skriftlighed som følge af prøveændringerne. Omvendt angiver 18 %, at de er uenige i, at de har arbejdet mere med skriftlighed. Det er der i interviewmaterialet flere mulige forklaringer på. Fx er der eksempler på lærere, der allerede arbejdede meget med skriftlighed. Men der er også lærere, der pointerer, at det er svært for dem at integrere skriftligheden i undervisningen. Der

er især to årsager, der går igen som forklaring herpå. For det første er det ifølge lærerne svært at integrere skriftlighed, fordi der ikke tidligere har været tradition for det i naturfagene. Det er nyt og svært for lærerne at skulle tænke i opgaver, der træner elevernes skriftlighed, for de har ikke været vant til det. Derfor er det nemt at gå tilbage til det, "man plejer at gøre" – særligt hvis man er presset på tid:

> *Men så kom jeg lidt fra det igen. Jamen, det der med, at de skulle lave en skriftlighed, når de laver noget, nogle forsøg. Altså have et hæfte ved siden af, ikk', og få skrevet lidt ned. Men igen, du ved, så kommer man hurtigt tilbage til sin egen undervisning, altså til den … altså til ens gamle undervisningsmåde, ikk'. […] Så det kræver, at man – at jeg kontinuerligt bliver holdt ved ilden eller ikke ved ilden, men du ved, bliver mindet om.* (Lærer)

Derfor savner flere lærere også konkrete skriftlige opgaver. Det uddyber vi i Kapitel 5 om barrierer.

Den anden forklaring på, hvorfor det er svært at integrere skriftlighed, er tidspres. Flere lærere oplever, at det er svært at prioritere skriftlighed, når de i naturfagene i forvejen er pressede og skal nå meget. Det vender vi tilbage til i Kapitel 5 om barrierer.

> *Jeg arbejder nærmest ikke med skriftlighed. […] Jeg tror i virkeligheden, at jeg sidder og tænker, jamen hvornår skulle de have tid til det.* (Lærer)

En enkelt lærer advarer mod, at undervisningen kommer til at fokusere for meget på skriftlighed. Læreren understreger vigtigheden af at lære eleverne, at naturfag ikke er dansk, og dermed at stavning og grammatik er mindre vigtigt i naturfag. Læreren anerkender, at skriftlighed vil fylde mere i naturfagsundervisningen, hvis forsøgsprøverne gøres permanente, og at det er godt, at elevernes naturfaglige argumentation styrkes, men stiller spørgsmålstegn ved, om skriftligheden er nødvendig:

> *Jeg tror egentlig bare, at jeg netop sætter spørgsmålstegn ved … hvornår det er nødvendigt, at de kan formulere sig skriftligt. Argumentere naturfagligt på skrift. Udover til sådan en prøve kan jeg ikke se, at det er nødvendigt, at de skal kunne argumentere naturfagligt på skrift.* (Lærer)

Så der kan altså findes eksempler på en vis skepsis over for skriftligheden og en utryghed ved at skulle gå den vej. De fleste lærere hilser dog skriftligheden velkommen og fremhæver, at det er en vej til at arbejde endnu mere med elevernes kommunikationskompetence. En lærer kalder skriftlighed i naturfagene

en forsømt disciplin og sammenligner situationen med det øgede fokus på faglig læsning i naturfagene.

> *Jeg tænker da, at den skriftlige dimension er fin at have med i naturfag, og vi har helt klart forsømt den. […] At det faktisk har været en totalt forsømt disciplin. Det var måske lidt ligesom dengang, vi lærte, at vi også skulle til at læse, at vi ikke kunne tage det som givet, at eleverne kunne læse teksterne i naturfag. Og selvfølgelig kan de ikke skrive i naturfag, for vi har jo ikke lært dem det, så hvorfor pokker skulle de kunne. Det tror jeg ikke, vi har haft fokus på.*
>
> **Lærer**

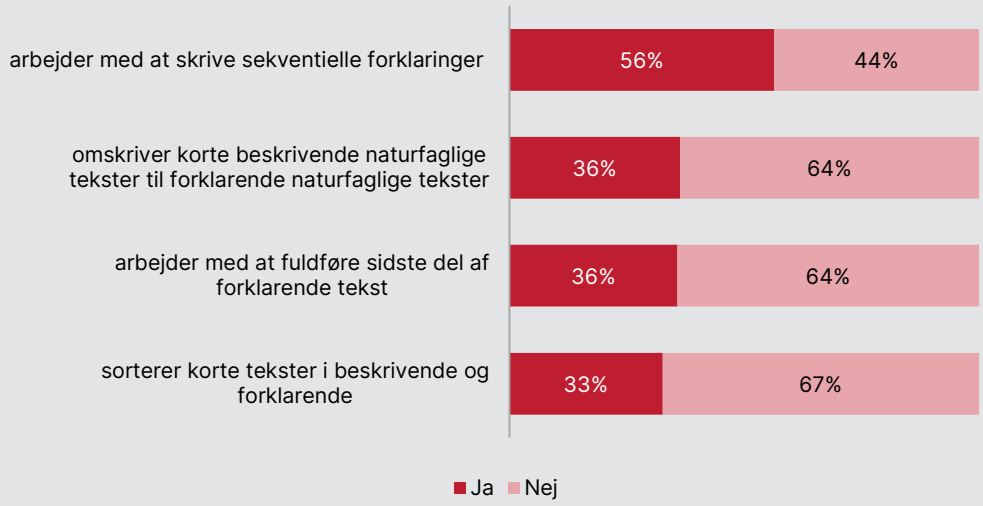
4.2.2 Fokus på argumentation, forklaring og fagbegreber

Lærerne ser altså overordnet et behov for at arbejde med elevernes skriftlighed i undervisningen. Mere specifikt ser de et behov for at arbejde med, hvordan man argumenterer, forklarer og bygger et svar op, der indeholder fagbegreber. Netop argumentation og forklaring har været et fokus for flere lærere og vil også være det fremadrettet:

> *(…) og vi var godt forberedte på, at vores undervisning den skule lægge mere op til det her med, at de skulle være mere forklarende og argumenterende i deres, i deres svar, når det var både mundtligt og selvfølgelig også skriftligt, ikk'. Men specielt det her med, når de, når vi laver noget undervisning, hvor nogle gange så skal de ud og undersøge et eller andet, og så skal de svare på nogle arbejdsspørgsmål, vi har lavet. Men der har vi opmærksomhed på, at vi stiller spørgsmål på en måde, der lægger op til, at der skal forklares og argumenteres.* (Lærer)

Pointen understreges af, at 56 % af lærerne har svaret ja til, at eleverne i løbet af skoleåret op til forsøgsprøverne har arbejdet med at skrive sekventielle forklaringer. Der er dog færre, der har arbejdet med forklaringer på de specifikke måder, der er spurgt ind til i de øvrige udsagn i Figur 4.3.

Figur 4.3 Hvordan der arbejdes med skriftlighed



Anm.: n = 133-138.

Spørgsmålsformulering: 'Har du gjort brug af følgende undervisningsaktiviteter det sidste skoleår? Eleverne …' 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

Nogle lærere fremhæver dog specifikt arbejdet med at kende forskel på beskrivelser og forklaringer som et fokus samt ikke mindst at træne eleverne i at genkende, hvornår henholdsvis en beskrivelse eller en forklaring efterspørges i en spørgsmålsformulering. Som det fremgår af figur 4.2, svarer 43 % af lærerne da også, at de som følge af den nye prøveform har arbejdet mere med signalord[8].

> *Jeg har også en del tosprogselever, så vi prøvede sådan at kigge på formuleringer, som er særligt geografi- og naturfagsformuleringer, og ligesom, hvad betyder det, når der står her, hvad er det så I skal. Og gjorde meget ud af at sige, jamen I skal skrive, hvad det er… Der er ikke et rigtigt facit, men der er mange rigtige facits i det, og det er jeres overvejelser, som er det vigtige i det her.*
> (Lærer)

Et andet fokus for flere lærere har været at bruge og forklare fagbegreber, så alle er med på, hvad de betyder:

[8] Ord, som sender signaler om, hvilke former for svar der forventes, eller hvad der skal til for at løse opgaven (Emu, 2023).

Så er jeg blevet bedre til at bringe fagord ind i – altså forklare fagord. (Lærer)

Endelig har et fokus været, hvordan et godt svar er bygget op. Det gode svar ser ikke anderledes ud skriftligt sammenlignet med, hvad eleverne er vant til mundtligt. Men der er stor forskel på at formulere sig mundtligt og skriftligt for eleverne:

Jeg var overrasket over så stor forskel, der er på at tale – snakke med en elev om det naturfaglige og så se deres svar på skrift, ikk'. Altså fordi når man taler med elever, så kan man hurtigt få en ide om, altså, om de har forståelse for faget, ikk', hvad der bliver formuleret i nogle sætninger, så kan det være sværere at vurdere, du ved, har jo egentlig forstået det, eller er det bare vrøvl, du skriver, ikk'. (Lærer)

4.2.3 Skriftlighed øves især i forbindelse med undersøgelser og forsøg

Lærerne vurderer det oplagt at træne skriftlighed i forbindelse med undersøgelser og forsøg. Flere lærere nævner det som et konkret eksempel på, hvordan de arbejder med skriftlighed i undervisningen. Flere peger på, at det er den måde, de arbejder på nu, og andre peger på, at det er oplagt at gøre fremadrettet. Konkret bruger lærerne forsøg som anledning til at træne eleverne i formulering af hypoteser, beskrivelse af forsøget og i at lave konklusioner, der følger op på hypoteserne.

> *Medmindre hypotesen er givet af mig, så skal de formulere en hypotese, og det taler vi jo så selvfølgelig om næsten hver gang, hvordan er det, vi gør det på en god måde. Og så skal de som regel også formulere en eller anden form for konklusion, når de har deres data fra en undersøgelse. Men som hovedregel handler det for mig om at lave mange små tekstøvelser.*
>
> **Lærer**

4.3 Simuleringers betydning for undervisningen

At simuleringer er integreret som en fast del af forsøgsprøverne ser også ud til at have betydning for undervisningen. Størstedelen af lærerne ser følgeligt et behov for at anvende simuleringer i undervisningen, men oplever, at udbuddet er småt – det vender vi tilbage til i Kapitel 5. Når simuleringer indgår i undervisningen, har eleverne typisk stor frihed i arbejdet med dem.

4.3.1 Lærerne ser gode muligheder i simuleringer og er i færd med at integrere dem

Som allerede nævnt finder vi i spørgeskemaundersøgelsen, at en af de største følger af prøveændringerne er, at lærerne i højere grad har arbejdet med simuleringer i undervisningen (se afsnit 4.1).[9] Vi ser en lignende tendens i interviewmaterialet. Alle interviewede lærere har været orienteret mod og bevidste om, at simuleringer skal være en del af undervisningen:

> *Ja, det har fyldt – vi havde mere fokus på at prøve at få noget med simuleringer med ind i undervisningen og sige til eleverne, de her forskellige hjemmesider, der er ret gode til at vise, hvad en simulering kan gøre for forståelsen.* (Lærer)

Det varierer imidlertid, i hvor høj grad simuleringer er blevet integreret i undervisningen. For nogle lærere har det været svært at få simuleringer ind i deres undervisning, både fordi det har været en udfordring at finde simuleringer og at finde ud af, hvordan man kan bruge simuleringer:

> *Men det har min fagteam-kollegaer fx. Der var en del af dem, der faktisk ikke rigtigt vidste, hvad en simulering var. Så det har jo været en kæmpe opgave at finde ud af, hvad kan jeg bruge dem til? Hvad kan simuleringer, hvor finder jeg simuleringer, der kan noget i forhold til mit fag?* (Lærer)

Hos andre lærere var simuleringer allerede en integreret del af undervisningen før forsøget. For dem har den største ændring været, at de har brugt simuleringer mere:

> *[Simuleringer] har jeg også brugt en del før. Også specielt under corona-undervisningen, så var det noget, jeg brugte rigtig, rigtig meget. Altså finde forskellige online simuleringer og så lave nogle*

[9] Simuleringer er blandt andet en vej til at arbejde med elevernes modelleringskompetence og nævnes i fagenes læseplaner som et eksempel på en interaktiv model.

opgaver, hvor de skal bruge de simuleringer til at lave nogle undersøgelser og så svare på nogle ting der. Så det bruger jeg en del på. (Lærer)

Størstedelen af de interviewede lærere henviser til PhET Colorado som deres primære og ofte eneste kilde til simuleringer. Flere påpeger dog, at det relevante udvalg af simuleringer er begrænset. Det er en barriere, vi vender tilbage til i afsnit 5.3.

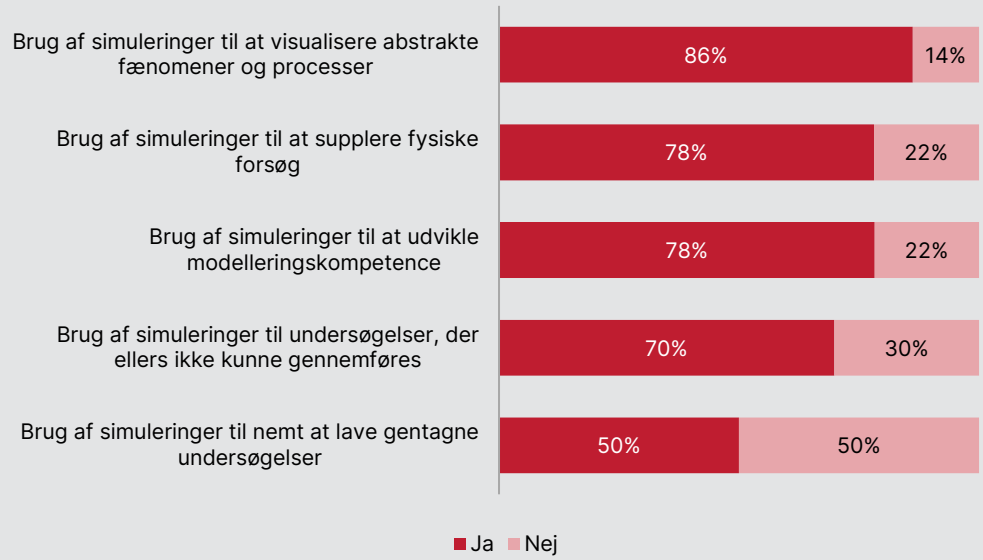
4.3.2 Eleverne gives typisk stor frihed i arbejdet med simuleringer

Interviewmaterialet indikerer, at simuleringer indgår i undervisningen på en måde, hvor arbejdet bliver mindre lærerstyret og mere elevstyret. Med det mener vi, at lærerne giver udtryk for, at eleverne ofte har stor frihed, når de arbejder undersøgende med simuleringer. Lærerne lader i høj grad eleverne prøve sig frem og gå undersøgende til værks:

De har sådan en atombygge-en, det kan være noget af det, jeg viser nogle elever som noget af det første. Så siger jeg, prøv lige at lege med det her, prøv lige – hvad er det I sidder med, tror I? Hvad foregår der, hvorfor kan I gøre det der, hvorfor virker det der ikke? Og stiller lidt nysgerrige spørgsmål ind til det. Og det kan også være noget, man bruger under et forløb, hvor man rent faktisk beder eleverne om – de har også en i forhold til kredsløb og universet – hvor man kan sige, der bliver den brugt mere aktivt i forhold til at få dem til at arbejde med kræfterne. Men den bruger vi mere, når vi allerede ved noget. Så det er meget forskelligt, men fælles for alle de forløb, hvor jeg bruger simuleringer, er, at jeg prøver at være meget lidt styrende på, hvor eleverne skal trykke henne. (Lærer)

Samtidig indikerer resultater af spørgeskemaundersøgelsen, der fremgår af Figur 4.4, at lærerne oftest anvender simuleringer til at visualisere abstrakte fænomener og processer – noget som situationerne beskrevet i citatet ovenfor er gode eksempler på. Simuleringer bruges også ofte som supplement til fysiske forsøg eller i tilfælde, hvor undersøgelsen ellers ikke kunne gennemføres, samt til at udvikle elevernes modelleringskompetence. Simuleringer anvendes sjældnere til nemt at lave gentagne undersøgelser.

Figur 4.4 Hvordan simuleringer anvendes



Anm.: n = 136-138.

Spørgsmålsformulering: 'Simuleringer kan anvendes på mange måder i undervisningen. Har du anvendt simuleringer på følgende måder?' 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

4.4 Meget begrænset betydning for eleverne

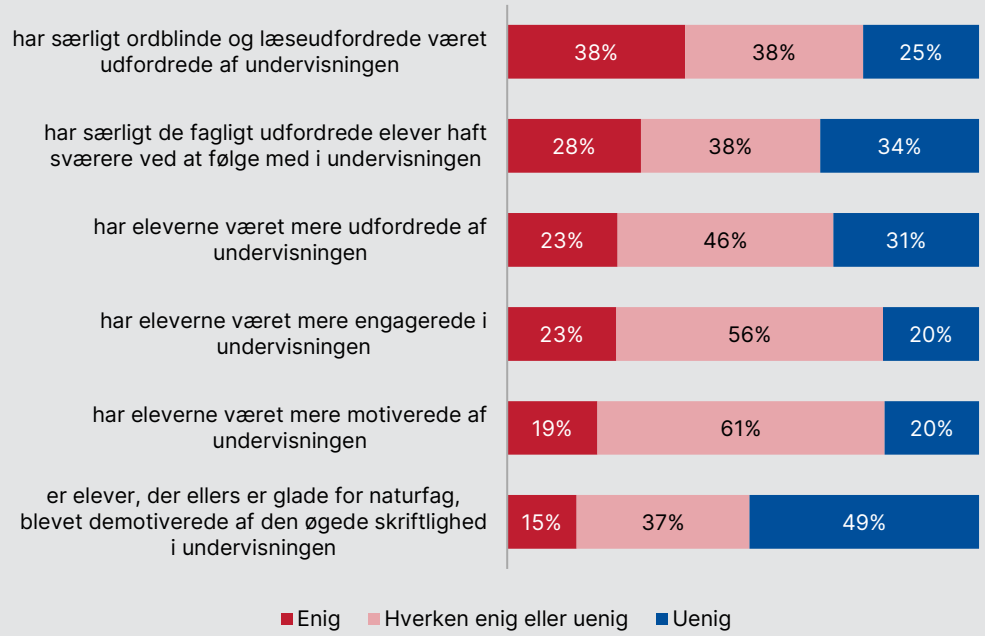
Vi har som en del af spørgeskemaundersøgelsen bedt lærerne vurdere, hvilken betydning prøveændringerne har for eleverne. Umiddelbart antyder lærernes svar, at betydningen for eleverne er meget begrænset. Resultaterne fremgår af Figur 4.5.

Ser man nærmere på lærernes svar, er den hyppigst anvendte svarkategori på næsten alle spørgsmålene 'hverken enig eller uenig'. Samtidig er der stort set balance mellem, hvor store andele der er henholdsvis enige og uenige i de undersøgte udsagn. Det gælder fx i forhold til elevernes engagement og motivation. Vi tolker det sådan, at lærerne ikke mener, at prøveændringerne har nogen særlig betydning for disse faktorer. Samme indtryk får vi, når vi ser på interviewmaterialet.

Lidt klarere holdninger kommer til udtryk, når lærerne skal forholde sig til, om prøveændringen betyder, at særligt ordblinde og læseudfordrede har været

særligt udfordrede af undervisningen. Det er 38 % nemlig enige i, mens en mindre andel – 25 % – er uenige. Det kommer også til udtryk under interviews, og det er ikke overraskende skriftligheden i forsøgsprøverne, der fremhæves som en udfordring. Som allerede nævnt i afsnit 2.3.1 ser elevernes danskfærdigheder dog ikke ud til at have større betydning for, hvordan de klarer sig i forsøgsprøverne end i de eksisterende prøver.

Figur 4.5 Betydning for eleverne



Anm.: n = 107-114.

Spørgsmålsformulering: 'Hvor enig eller uenig er du i følgende udsagn? Som følge af den nye prøveform …'. Enig = 'enig' eller 'meget enig'; uenig = 'uenig' eller 'meget uenig'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

4.5 Uændret fokus på den fælles prøve, men tegn på bedre sammenhæng

Et væsentligt opmærksomhedspunkt omkring de naturfaglige udtrækprøver er, at de ikke står alene, men udgør et supplement til den fælles praktisk-mundtlige naturfagsprøve. Sammenligner man de to prøveformer, har der væ-

ret stor forskel på dem, både hvad angår form og indhold. Men med ændringerne i forsøgsprøverne har man blandt andet ønsket at skabe en tættere sammenhæng mellem de to prøver, så udtrækprøverne i højere grad også afprøver elevernes naturfaglige kompetencer ligesom den fælles prøve. Vores analyser viser umiddelbart tegn på, at det er lykkedes.

Som nævnt tidligere har lærerne en klar opfattelse af, at forsøgsprøverne i højere grad end de eksisterende prøver afprøver elevernes naturfaglige kompetencer. Lærerne har dog fortsat fokus på at forberede eleverne til den fælles naturfagsprøve, mens det eksplicitte fokus på udtrækprøverne – her forsøgsprøverne – træder mere i baggrunden for samtlige interviewede lærere. Den fælles prøve er obligatorisk, og de fællesfaglige forløb fylder meget i den naturfaglige undervisning. Det er forberedelsen af eleverne til den mundtlige prøve, der prioriteres, og det har ikke ændret sig med forsøgsprøverne.

> *Der ligger der klart mest vægt på den fælles prøve. Og så... jamen det gør der jo, og hvis vi så bliver trukket, så må vi jo lige lave lidt intensiv kursus til sidst. Ikke, at vi ikke er omkring tingene, men det er ikke der, hvor vægtningen ligger, for vi ved jo ikke, om vi bliver trukket.*
>
> **Lærer**

Men i kraft af at forsøgsprøverne passer bedre sammen med den fælles prøve, er der blandt de interviewede lærere en oplevelse af, at undervisningen, der er orienteret mod den fælles prøve også forbereder eleverne på forsøgsprøverne. Der er dermed større balance mellem forberedelsen til de monofaglige skriftlige prøver og den fælles prøve i undervisningen, på trods af at undervisningen ikke har ændret fokus.

> *Jeg følte, at min undervisning skulle være delt rigtig meget op, så eleverne vidste, jamen det her det er fællesfagligt. Der blander vi fagene her, det her, det er ren fysik/kemi. Jeg synes, de har svært ved at, jamen hvorfor kan vi ikke have noget af det andet med i det også. De har svært ved at forstå, at det skulle være det der fagspecifikke. Men her, der synes jeg måske lidt, sådan lige i måden, det er bygget op på prøven der, så minder det måske lidt om det med de simuleringer og muligheden for de kompetencer at få dem i spil. Så der har vi da snakket om, at det må være lidt nemmere at få lagt undervisningen om.* (Lærer)

Spørgeskemaundersøgelsen antyder dog en lidt mere balanceret holdning hos lærerne. Direkte adspurgt svarer 38 % af lærerne således, at de som følge af prøveændringerne har bedre mulighed for både at forberede eleverne til den fælles prøve og udtræksprøverne. Det er 24 % af lærerne omvendt uenige i (se Figur 4.2).

5 Oplevelse af forsøget og barrierer

Som en del af evalueringen har vi også undersøgt, hvordan lærerne oplever at tage del i forsøget og ikke mindst, hvilke barrierer for forandring lærerne oplever. Konkret har vi taget udgangspunkt i følgende undersøgelsesspørgsmål:

- Har lærerne været tilstrækkeligt klædt på til at tilrettelægge undervisningen frem mod prøverne?
- Hvilke barrierer har lærerne oplevet i forbindelse med tilrettelæggelsen af undervisningen frem mod prøverne?

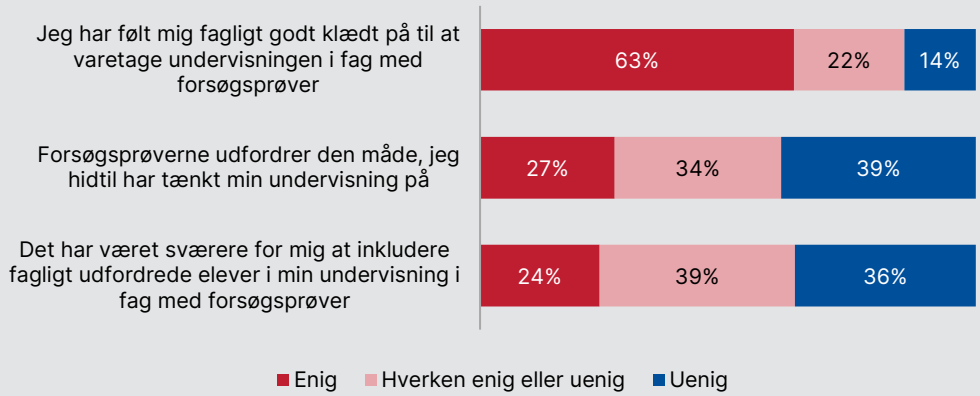
Igen baserer analyserne i kapitlet sig på interviewdata og spørgeskemaundersøgelsen blandt lærere. De samlede interviewdata er inddraget i analysen, men også i dette kapitel har hovedfokus været på lærerinterviewene gennemført i efteråret 2023, da disse bidrager med de mest opdaterede perspektiver på forsøget og barrierer.

Vi finder, at der er god stemning omkring forsøget. Lærerne er glade for at deltage og føler sig fornuftigt klædt på til det. Men der kan opstå flaskehalse i kommunikationen fra ministeriet, og der er en stor efterspørgsel på konkret undervisningsmateriale i form af simuleringer, skriftlige opgaver og eksempler på prøveopgaver. Implementeringen af skriftlighed og simuleringer sinkes desuden af begrænset forberedelsestid.

5.1 Overordnet god stemning omkring forsøget

Lærerne har overordnet følt sig godt fagligt klædt på til at tilrettelægge undervisningen frem mod prøverne. Det giver 63 % af lærerne udtryk for i spørgeskemaundersøgelsen. Blot 14 % giver udtryk for det modsatte (se Figur 5.1). Samme tendens er tydelig i interviewmaterialet, hvor det fremhæves, at kollegial sparring i fagteams har været central for netop følelsen af at føle sig klædt på.

Figur 5.1 Potentielle udfordringer



Anm.: n = 137-140.

Spørgsmålsformulering: 'Hvor enig eller uenig er du i følgende udsagn?' Enig = 'enig' eller 'meget enig'; uenig = 'uenig' eller 'meget uenig'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

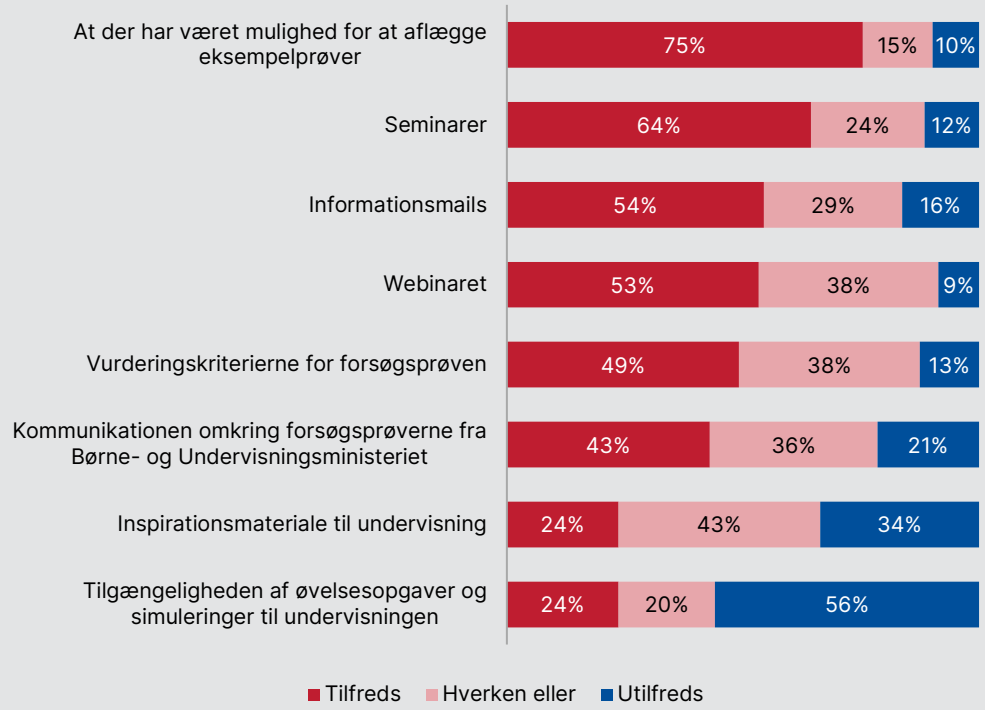
Generelt er der også en god stemning omkring forsøget. Lærerne er glade for at være med og støtter op om ændringerne. Lærernes overordnede tilfredshed kommer også til udtryk i spørgeskemaresultaterne, der fremgår af Figur 5.2. Fx er langt størstedelen af lærerne godt tilfredse med, at der har været mulighed for at aflægge eksempelprøver og at deltage i seminarer.

Flere lærere giver også udtryk for, at deres deltagelse i forsøget inspirerer dem til at arbejde med deres fag. En lærer giver eksempelvis udtryk for, at prøveændringerne kan være med til at skubbe til, hvordan nogle lærere tænker naturfagene og undervisning – og at det er en god ting.

> *Det tvinger os alle sammen til at tænke faget på en anden måde, og det er også godt for sådan en lærer, der siger, mit fag – min undervisning. Ja, det er godt med dig. Der kan vi blive skubbet til på en god måde.*
>
> **Lærer**

Men lærerne oplever samtidig en række barrierer, som vi uddyber nedenfor.

Figur 5.2 Tilfredshed med forsøget



Anm.: n = 104-127.

Spørgsmålsformulering: 'Hvor tilfreds har du været med...?' Tilfreds = 'tilfreds' eller 'meget tilfreds'; utilfreds = 'utilfreds' eller 'meget utilfreds'. 'Ved ikke'-svar er sorteret fra.

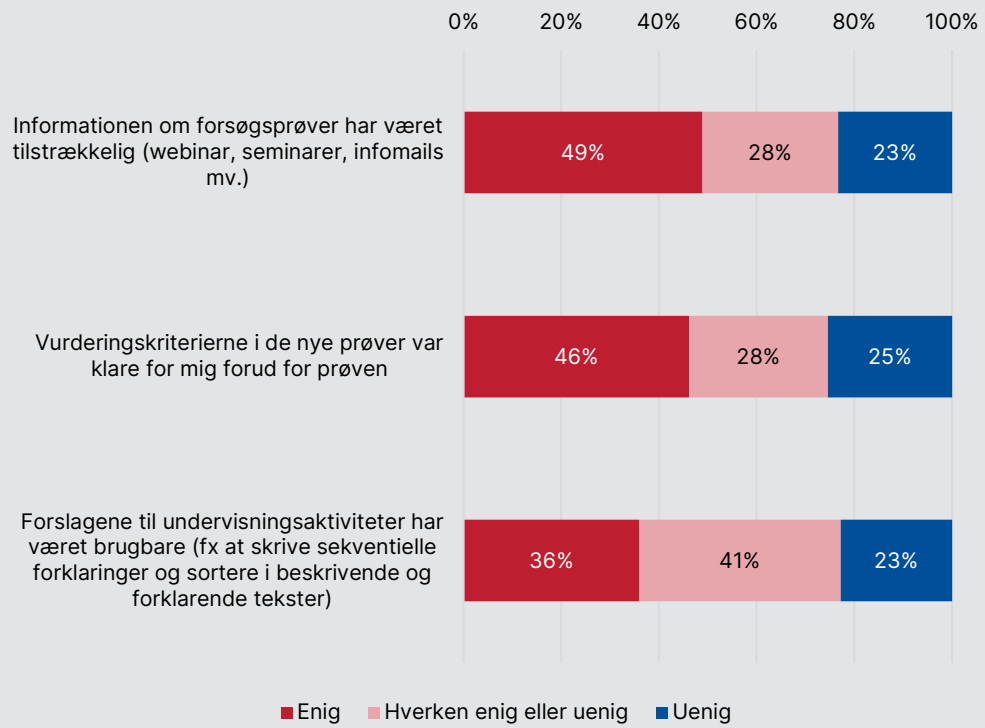
Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

5.2 Risiko for flaskehalse

Som det eksempelvis indikeres af lærernes tilfredshed med kommunikationen omkring forsøgsprøverne fra Børne- og Undervisningsministeriet, har lærerne til tider oplevet udfordringer med kommunikationen om forsøget. Lærere uddyber under interviews, at informationen egentlig har været klar nok, men at den ikke altid er nået frem til de rigtige lærere i tide. Det kan fx være i tilfælde, hvor en leder er kontaktperson. Det giver måske den fordel, at skolens deltagelse i forsøget legitimeres over for lærerne, men lederen kan samtidig blive en flaskehals for informationen til lærerne.

Jeg tror netop det der med, at fordi jeg ikke har været kontaktperson, så har det været sådan lidt, at det er kommet sådan lidt … drypvis meget, hvor så er min leder kommet og sagt: "Kom, nu skal I gøre det her", og så har jeg sagt: "Okay, det ved jeg ikke, hvad jeg skal gøre ved". Jeg ved det stadig ikke. Og så har jeg fået sendt en masse links og ting, som jeg skulle læse og sådan noget. Så på den måde. (Lærer)

Figur 5.3 Vurdering af kommunikation



Anm.: n = 114-133.

Spørgsmålsformulering: 'Hvor enig eller uenig er du i følgende udsagn?' Enig = 'enig' eller 'meget enig'; uenig = 'uenig' eller 'meget uenig'. 'Ved ikke'-svar er sorteret fra.

Kilde: VIVE-spørgeskemaundersøgelse til lærere, 2023.

5.3 Mangel på konkret materiale

Den barriere, der fylder mest blandt lærerne, er dog en generel oplevelse af mangel på konkret undervisningsmateriale. Som det fremgår af Figur 5.3, vur-

derer lærerne egentlig overordnet, at ministeriets forslag til undervisningsaktiviteter er brugbare, men når de vurderer tilfredsheden med det tilgængelige inspirationsmateriale (se Figur 5.2), er billedet mere negativt. I interviewmaterialet er der flere forklaringer herpå.

Først og fremmest savner lærerne i høj grad **simuleringer**. Lærerne oplever, at udbuddet er for småt og særligt, at der mangler simuleringer på dansk. De fleste lærere kender udelukkende til simuleringerne fra PhET Colorado[10]. Det begrænser lærernes muligheder og gør, at implementeringen af simuleringer i undervisning ikke går så stærkt, som den kunne gøre. Lærerne vil gerne bruge simuleringerne, men de er bare ikke altid tilgængelige.

> *Det er jo stadigvæk en kæmpe udfordring, fordi vi har for lidt* [simuleringer]. *Der kommer mere, jeg kan godt se, der begynder at poppe lidt forskelligt op rundt omkring. Men vi har jo ikke den der store base med fede simuleringer, vi lige kan bruge. Og det er et problem, kan man sige.*
>
> **Lærer**

Lærerne oplever, at der særligt mangler simuleringer omkring geografi.

Lærerne efterspørger også konkrete **skriftlige opgaver**, der matcher kravene i forsøgsprøverne – eller som minimum et større inspirationsmateriale til, hvordan de selv kan lave de nødvendige tilpasninger af deres undervisning.

> *Jamen ja, inspirationsmateriale er i virkeligheden et godt ord* [om hvad der mangler]. *Men man kan sige, denne her bank af forskellige ting, der kan hjælpe, når jeg nu vil træne mine elever i, hvordan er det nu lige, jeg laver en god hypotese, det er der altså ikke. Der er rigtig mange, der skriver: "lav en hypotese til det her". Der er også rigtig mange steder, man kan finde hypoteser. Men hvordan ser den gode hypotese ud? Astra kan meget og gør også en hel del her, men det kunne jo være fedt, hvis det også var noget, der var integreret lidt bedre i de systemer, vi alle sammen arbejder med. Fordi der er jo ikke nogen af os, der ikke arbejder med et system.* (Lærer)

Flere lærere giver under interviews udtryk for, at deres undervisning i høj grad baserer sig på, hvad end der er af indhold på de undervisningsportaler, de har

[10] (PhET, 2023) som allerede var en offentligt tilgængelig ressource inden forsøget.

adgang til. Og her er der ifølge lærerne typisk mangel på både simuleringer og skriftlige opgaver. Lærerne anerkender, at der vil være et efterslæb, fordi prøveændringerne stadig er på forsøgsstadiet, men de har samtidig en klar forventning om, at både forlagene og ministeriet stiller mere materiale til rådighed, hvis prøveændringerne permanentgøres.

Sidst, men ikke mindst, efterspørger lærerne flere **eksempler på, hvordan forsøgsprøverne kan se ud**. Der har været gennemførte to eksempelprøver, og det er godt, men lærerne vil gerne for både deres egen og elevernes skyld have flere konkrete eksempler. Det vil hjælpe dem med at forberede eleverne, så det er klart for dem, hvad de kan forvente, og de kan gå mindre usikre ind til prøverne.

> *Der kom en frustration oveni hos en del kollegaer, fordi de jo så netop heller ikke selv kunne sætte sig ned og se, hvad er det så, eleverne skal. De manglede også selv de her eksempler på, jamen hvad er det, der skal foregå. Så det har været svært, og det har også været svært at prøve at hjælpe og støtte med, for jeg har jo heller ikke haft eksemplerne. Så det har jo været sådan lidt på, nå men hvis man nu gør sådan her, og de her opgaver, vi plejer at bruge her, det kunne man jo godt … Men for nogle, mere end andre, har det føltes lidt som at famle rundt i blinde. Og skulle lede nogle elever afsted mod noget, man faktisk ikke rigtigt vidste, hvor lå henne.* (Lærer)

Ønsket om mere konkret materiale understreges af, at 56 % af lærerne i spørgeskemaundersøgelsen udtrykker utilfredshed med tilgængeligheden af øvelsesopgaver og simuleringer til undervisningen (se Figur 5.2).

5.4 Tid som en knap ressource

Lærerne oplever også en praktisk udfordring i, at deres tid er meget begrænset. Særligt i biologi og geografi har lærerne ganske få undervisningstimer og dermed meget lidt forberedelsestid. Det gør det svært for dem at implementere de nødvendige tilpasninger af deres undervisning og at fx udvikle egne undervisningsmaterialer.

> *Du har en time om ugen i både geografi og biologi, samtidig med at du skal igennem de her fælles faglige fokusområder, som de skal op i til den fælles prøve. Det levner ikke ret meget tid til sådan nogle ting. […] Men det levner ikke ret meget tid til de der fagfaglige ting,*

> *som vi prøver på at putte ind i det fælles faglige. Så derfor bliver tiden til sådan noget som skriveøvelser og simuleringer, det bliver svært at finde.* (Lærer)

Fordi fagene er "små", ligger de også ofte yderst i skemaerne, hvor de er ekstra sårbare for aflysninger. Derfor holder lærerne sig typisk meget tæt til deres årsplan. Herudover har selve deltagelsen i forsøget også taget tid.

Oplevelsen af at mangle tid til at implementere de nødvendige ændringer kan evt. også bidrage til at forklare, at 27 % af lærerne oplever, at forsøgsprøverne udfordrer den måde, de hidtil har tænkt deres undervisning på (se Figur 5.1). Vores analyser tyder dog på, at der også er andre faktorer i spil. Vi finder eksempelvis, at de erfarne lærere i højere grad end de mindre erfarne lærere oplever forsøgsprøverne som en udfordring. Netop de erfarne lærere vil typisk over en årrække have udviklet opgaver og en tilgang til deres undervisning, som de bygger videre på år efter år. Uagtet forsøget vil de mindre erfarne lærere i højere grad have behov for at udvikle deres tilgang, og prøveændringen vil derfor ikke nødvendigvis opleves som noget, der kræver ekstra tilpasning.

6 Konklusion

Samlet set finder vi, at forsøgsprøverne har en passende sværhedsgrad, og at der på tværs af alle tre forsøgsprøver og både del A og B er en acceptabel grad af konsistens. Dette indikerer, at prøvernes opgavedele er sammenhængende og samlet set giver et mål for dygtighed på en underliggende dimension. Imidlertid opstår der udfordringer, når det kommer til endimensionalitet, lokal uafhængighed og stabilitet. Disse aspekter tyder på, at prøverne potentielt kan generere mere støjfulde resultater end det, der er ideelt set ud fra et testteoretisk perspektiv. Det er dog vigtigt at bemærke, at Rasch-modellen udgør en ret restriktiv vurderingsramme, og de nævnte udfordringer bør overvejes i konteksten af et behov for at ramme en balance mellem at opretholde robuste testteoretiske egenskaber og at evaluere elevernes dygtighed inden for brede kompetenceområder ved hjælp af forskellige metoder. Specifikt for situationer, hvor prøverne er adaptive, og eleverne kun modtager et udsnit af spørgsmålene (som for eksempel ved de nationale test), eller hvor prøverne gentages årligt, kan overtrædelser af Rasch-modellens antagelser være særligt problematiske. I sådanne tilfælde kan estimater af elevernes dygtighed variere markant på grund af eksponeringen for forskellige items med varierende stabilitet og afhængighed. Men i dette tilfælde, hvor alle elever modtager alle spørgsmål, og spørgsmålene er udviklet specifikt til hver kohorte, er denne bekymring mindre kritisk.

Evalueringen viser således, at mens forsøgsprøverne opfylder visse centrale kvalitetskrav, er der potentielt plads til forbedring, især med hensyn til at sikre større stabilitet og uafhængighed mellem opgaverne. Det er vigtigt at fortsætte med at overvåge og justere prøverne for at forbedre deres pålidelighed og validitet, særligt i lyset af de udfordringer, der er identificeret i forhold til endimensionalitet, lokal uafhængighed og stabilitet. Dette vil hjælpe med at sikre, at prøverne forbliver et retfærdigt og nøjagtigt mål for elevernes dygtighed.

Udfordringerne med Rasch-modellens grundantagelser er dog ikke større i prøvernes nye del sammenlignet med den eksisterende del, og vores analyser indikerer derfor ikke, at tilføjelsen af blandt andet korte tekstsvar og interaktive simuleringer gør prøveegenskaberne dårligere.

Vi finder heller ikke problemer i, at ændringerne medfører behov for manuelle vurderinger af prøvebesvarelserne. Censorerne er gode til at følge rettevejledningerne, og der er acceptabel enighed mellem dem. Samtidig viser vores øvrige analyser, at både lærere og elever grundlæggende ser positivt på ændringerne, og at lærerne mener, at man med forsøgsprøverne er gået et skridt i den rigtige retning. Der er bekymringer omkring, om skriftligheden i forsøgsprøverne vil have konsekvenser for de sprogligt udfordrede elever og stille

dem dårligere end ved de eksisterende prøver, men vi finder ingen tegn på, at det faktisk er tilfældet i de statistiske analyser. Herudover er der dog blandt lærerne en ret udbredt efterspørgsel på, at de understøttes bedre i at lave de nødvendige ændringer af deres undervisning – konkret gennem mere undervisningsmateriale – men lærerne er allerede godt i gang på trods af en oplevelse af mangel på tid.

Alt i alt er vores konklusion på den baggrund, at der hverken er noget ved selve ændringerne i forsøgsprøverne eller i oplevelsen af dem og deres konsekvenser, der står i vejen for, at ændringerne gøres permanente. Det er dog væsentligt fortsat at overvåge og justere prøverne, arbejde med kommunikationen til skolerne og at sikre, at lærerne støttes tilstrækkeligt i implementeringsprocessen.

DEL 2
Dokumentation

7 Data og metoder

I dette kapitel dokumenterer vi vores arbejde gennem uddybende beskrivelser af evalueringens datagrundlag og de analysetilgange, vi har anvendt.

7.1 Prøve- og registerdata

De kvantitative analyser af forsøgsprøverne laves på baggrund af en kombination af registerdata fra Danmarks Statistik og selve prøvedataene fra biologi, geografi og fysik/kemi samt data om resultater fra en række andre prøver ved folkeskolens afgangseksamen fra Styrelsen for Undervisning og Kvalitet (STUK).

7.1.1 Prøvedata fra STUK

Prøvedataene består af to datasæt, ét fra henholdsvis del A og del B, som hver især indeholder alle resultaterne for alle tre fag. Begge datasæt indeholder oplysninger om, hvilken score den enkelte elev har fået i hvilket item i hvilket fag. Data fra prøvernes del B indeholder for hvert item en score givet af både censor 1 og censor 2. Derudover indeholder begge datasæt et institutionsnummer, så vi kan identificere, hvilken skole eleverne går på, og en pseudonymiseret form af elevernes CPR-nummer, der gør os i stand til at koble baggrundsoplysninger om eleverne på data. Ved prøveafviklingen opstod der en fejl, som betød, at en opgave måtte tages ud af biologiprøvens del B, og ingen elever har derfor fået point i denne opgave, hvorfor den ikke indgår i analyserne i denne undersøgelse. Der er ikke det samme antal elever, som har besvaret både del A og del B. Dette kan enten skyldes, at en skole har haft et internetnedbrud og derfor ikke har kunnet gennemføre den ene del online. Derudover er der flere elever på særlige vilkår, der laver del A i hånden, hvorfor disse svar ikke vil indgå. Om disse tre grunde er årsag til det forskellige antal besvarelser, kan vi imidlertid ikke være sikre på. Analyserne er derfor kun baseret på de elever, som har besvaret begge delprøver.

Dataene fra andre prøver ved folkeskolens afgangseksamen indeholder det antal point, eleverne har fået i dansk læsning, dansk retskrivning og matematik uden hjælpemidler. Dette datasæt indeholder ligesom prøvedataene et institutionsnummer og en pseudonymiseret form af elevernes CPR-nummer. Data fra de øvrige fag bliver brugt til at undersøge sammenhængen mellem resultaterne i forsøgsprøverne og elevernes dygtighed i andre dag. Vi har valgt disse

tre fag ud fra deres tilgængelighed og ud fra en hypotese om, at øget skriftlighed, målt ved resultaterne i dansk, kan have større betydning for den nye del af forsøgsprøverne.

7.1.2 Registerdata fra Danmarks Statistik

Fra Danmarks Statistiks registre har vi forskellige oplysningerne om eleverne og deres forældres baggrund. Det er oplysninger som oprindelsesland, køn, alder, forældrenes højest fuldførte uddannelse og forældrenes indkomst. Vi har også oplysninger om elevernes standpunktskarakterer i årene før folkeskolens afgangseksamen i sommeren 2023.

7.2 Lærerspørgeskema

Spørgeskemaet blandt lærerne i forsøgsprøvefagene havde til formål at undersøge, hvordan lærere oplever prøven og dens indflydelse på undervisningen, herunder brugen af simuleringer og arbejdet med skriftlighed i naturfagene. Det søger blandt andet at afdække, hvilke greb lærerne benyttede sig af i undervisningen frem mod forsøgsprøverne. Har underviserne følt sig klædt på til at forberede eleverne til prøven, og hvilke barrierer har de oplevet?

Spørgeskemaet er blevet udviklet i samarbejde med STUK og er desuden blevet pilottestet ved hjælp af cognitive interviewing-metoden (Beatty & Willis, 2007). Der er i den forbindelse udført to interviews med kontaktpersoner, som under interviewet blev bedt om at gennemgå spørgeskemaet og "tænke højt" imens. Pilottesten ledte til en række ændringer i spørgeskemaet, som både angik forståelse af spørgsmålene samt tilføjelser af spørgsmål.

Spørgeskemaet blev udsendt til kontaktpersonerne for forsøgsprøverne på forsøgsskolerne den 6. juni 2023, og undersøgelsen lukkede den 11. juli. Da hverken VIVE eller STUK har haft informationer om, hvem der har undervist i forsøgsprøvefagene frem mod prøven, har vi valgt at sende et link til spørgeskemaet til kontaktpersonerne på skolerne. Kontaktpersonerne blev i denne henvendelse bedt om at videresende linket til de relevante lærere. Dermed kender vi ikke populationens præcise størrelse og kan heller ikke udregne en svar%. Vi ved dog, hvilke skoler svarene kommer fra. Derfor ved vi også, at vi har mindst ét svar fra 90 % af de skoler, som har deltaget i forsøget. I alt 132 lærere gennemførte hele spørgeskemaet.

Efter udsendelsen den 6. juni er der blevet fulgt op med e-mailrykkere til kontaktpersonerne den 12. juni og den 20. juni. Desuden er der foretaget en rundringning til skoler uden svar i perioden 13. juni – 22. juni.

7.3 Interviewdata

Interviewdata består af fokusgruppeinterviews med lærere, fokusgruppeinterviews med elever og individuelle lærerinterviews. Fokusgruppeinterviewene med både lærere og elever havde eksempelprøverne som omdrejningspunkt. Eksempelprøverne blev gennemført ad to omgange: I oktober 2021 og december 2022.

De individuelle lærerinterviews havde forsøgsprøverne som omdrejningspunkt. Forsøgsprøverne blev gennemført i forbindelse med folkeskolens afgangsprøve i sommeren 2023. Interviewene har til formål at belyse, hvordan lærere og elever oplevede prøverne, og hvilken betydning prøveændringerne har for undervisningen. Vi beskriver de forskellige interviews nedenfor. Alle interviews er dokumenteret med enten fyldige referater eller transskriberinger. Interviewene er efterfølgende blevet systematisk og tematisk kodet ud fra de relevante undersøgelsesspørgsmål.

7.3.1 Lærerinterviews

Vi har gennemført videointerviews med ni lærere fra forskellige skoler, der har ført en klasse op til én af de tre forsøgsprøver ved folkeskolens afgangsprøver i sommeren 2023. Interviewene er gennemført med tre lærere fra hvert af de tre naturfag.

Lærerne er udvalgt gennem tilfældigt udtræk af ni klasser fra forskellige skoler, der blev udtrukket til én af de tre forsøgsprøver i sommeren 2023. Der er udvalgt tre klasser, der blev udtrukket i den skriftlige prøve i hhv. fysik/kemi, biologi og geografi. Klasserne er udtrukket tilfældigt, men med hensyntagen til geografisk placering. Således er de fleste landsdele repræsenteret, nemlig Nord-, Midt- og Sydjylland, Fyn, Nord- og Midtsjælland samt Københavnsområdet. Alle tre fag er ligeledes repræsenteret i forskellige landsdele, enten både Jylland og Sjælland eller både Fyn og Sjælland. Efter udtrækningen blev skolens kontaktperson kontaktet med henblik på at blive sat i forbindelse med den udtrukne klasses lærer i det pågældende naturfag. Fire klasser blev genudtrukket, da den pågældende lærer enten ikke længere var ansat på skolen eller ikke ønskede at deltage.

De interviewede lærere er fra skoler af varierende størrelse (mellem ca. 80 til 700 elever), med varierende elevgrundlag (både overvægt af svage og over-

vægt af stærke elever) og har varierende erfaring (4-35 års erfaring). Syv lærere er ansat på folkeskoler, og to lærere er ansat på frie grundskoler. 8 ud af 9 lærere underviser også i mindst ét andet naturfag end det, de har ført deres klasse til prøve i.

Interviewene er gennemført i perioden oktober-november 2023. Hvert interview var af ca. én times varighed og blev gennemført med udgangspunkt i en semistruktureret interviewguide ud fra følgende temaer:

- Oplevelse af forsøgsprøverne
- Forsøgsprøvernes betydning for undervisningen
- Forsøgsprøvernes betydning for eleverne
- Oplevelse af at være en del af forsøget.

Alle interviews er optaget efter indhentning af verbalt samtykke fra læreren og transskriberet. Interviewene er efterfølgende blevet systematisk og tematisk kodet på baggrund af de relevante undersøgelsesspørgsmål.

7.3.2 Fokusgruppeinterview med lærere og elever

Der er gennemført fokusgruppeinterview med lærere og elever ad to omgange. Den første runde af casebesøg blev gennemført i perioden februar-marts 2022 på seks skoler, der gennemførte eksempelprøverne i oktober 2021. Den anden runde blev gennemført ved casebesøg på seks skoler i forbindelse med eksempelprøverne i december 2022.

Fra runde 1:

Der blev gennemført fokusgruppeinterviews med i alt 13 lærere og 29 elever på seks skoler, som gennemførte eksempelprøverne i efteråret 2021. De seks skoler var udvalgt efter, hvorvidt de havde gennemført mindst to eksempelprøver i enten fysik/kemi, geografi eller biologi. Skolerne var derudover udvalgt med hensyn til geografisk placering, hvoraf tre skoler ligger i Jylland og tre skoler ligger på Sjælland. Hver prøve er således repræsenteret på to skoler – én i Jylland og én på Sjælland. Alle deltagende skoler gennemførte både del A og del B af eksempelprøverne.

Fokusgruppeinterviewet med eleverne havde primært fokus på elevernes oplevelse af eksempelprøverne. Derudover fokuserede interviewet på, hvordan eleverne oplever prøvernes sammenhæng med deres undervisning.

Fokusgruppeinterviewet med lærer havde primært fokus på lærernes forventninger til forsøgsprøverne, deres tanker om eksempelprøverne samt deres undervisningsmæssige udfordringer med prøverne.

Alle interviews er gennemført med udgangspunkt i semistrukturerede interviewguides. Referaterne af interviewene er kodet systematisk efter overordnede temaer.

Interviewene blev gennemført i vinteren 2022 i perioden 23. februar-21. marts. For at genopfriske både lærere og elevers hukommelse var print af eksempelprøverne tilgængelige under interviewene.

Fra runde 2:

Som i runde 1 besøgte vi seks udvalgte forsøgsskoler. Skolerne blev udvalgt, så der var to skoler, der havde gennemført hver af de tre prøver.

På hver af de seks skoler blev der gennemført to fokusgruppeinterviews med hhv.:

- 4-6 9. klasse-elever, der har gennemført eksempelprøven
- 2-4 lærere, der underviser i et naturfag på 9. årgang.

På tre af skolerne sad VIVE desuden med til eksempelprøven for at få et bedre udgangspunkt for at gennemføre de efterfølgende interviews.

8 Supplerende analyseresultater

Dette kapitel indeholder supplerende baggrundsanalyser af data fra forsøgsprøverne gennemført i sommeren 2023. Vi starter med at undersøge, hvem der har deltaget i forsøget, og derefter følger kapitlet samme struktur som Kapitel 2, således at vi først præsenterer detaljerede resultater vedrørende prøvernes sværhedsgrad efterfulgt af analyser af prøvernes grundlæggende egenskaber og forskelle på tværs af elevgrupper.

8.1 De deltagende elever i forsøget

Dette afsnit indeholder en beskrivelse af de elever, som deltager i forsøget. Vi undersøger, om de elever, der deltager i prøverne, er anderledes end elever, der ikke deltager i prøverne, og om forskellige elevgrupper er repræsenteret blandt de elever, som deltager i prøverne. Derudover undersøger vi, om skoler og elever i hele landet er repræsenteret.

Vi finder, at de deltagende elever er repræsentative for elever i 9. klasse.

I alt har 2.676 elever besvaret forsøgsprøvernes del A og 2.589 elever forsøgsprøvernes del B i biologi, geografi eller fysik/kemi. Der er således ikke lige mange elever, der har besvaret de to dele. Der er 424 elever, der har besvaret del A, men ikke del B, og 337 elever, der har besvaret del B, men ikke del A. 2.252 elever har besvaret begge prøver. 2.246 af disse kan vi koble registerdata på, og det er denne gruppe elever, alle analyserne af forsøgsprøverne er lavet på baggrund af.

De elever, som deltager i forsøgsprøverne, går på skoler, som har tilmeldt sig forsøget på eget initiativ. Der er derfor en risiko for, at de skoler, som har valgt at deltage i forsøget, og dermed deres elever, er forskellige fra andre skoler og deres elever. For at finde ud af om dette er tilfældet, sammenligner vi de elever, som deltager i forsøget, med alle andre elever i 9. klasse, som ikke deltager i forsøget. Det gør vi på en række socioøkonomiske karakteristika relateret til elevernes og deres forældres baggrund. Tabel 8.1 viser forskellene mellem de to grupper.

Tabel 8.1 Forskel mellem elever, der deltager i forsøget, og elever, som ikke gør

	Elever, som deltager i forsøget		Elever, som ikke deltager i forsøget		Forskel	P-værdi
Karakteristika	**Gennemsnit**	**Standard-afvigelse**	**Gennemsnit**	**Standard-afvigelse**		
Pige	0,51	0,50	0,49	0,50	0,02	**0,04**
Antal børn, far	1,12	0,35	1,13	0,35	-0,01	0,29
Antal børn, mor	1,12	0,34	1,13	0,35	-0,01	0,48
Dansk oprindelse	0,88	0,33	0,88	0,32	-0,01	0,45
Alder pr. 1. maj 2023	14,91	0,37	14,93	0,40	-0,03	**0,00**
Indkomst, far, DKK	395.723	340.283	383.425	444.109	12.298	0,21
Indkomst, mor, DKK	324.470	790.758	304.002	188.330	20.468	**0,00**
Antal års skolegang, far	13,68	2,65	13,57	2,61	0,11	0,06
Antal års skolegang, mor	14,00	2,59	13,92	2,70	0,08	0,20
Karaktergennemsnit 8. klasse	6,76	2,11	6,39	2,29	0,37	**0,00**
Gennemsnitlig indkomst, forældreparret, DKK	355.474	437.265	339.714	252.569	15.760	**0,01**
Gennemsnitlig indkomst over gennemsnit, forældreparret	0,54	0,50	0,51	0,50	0,03	**0,01**
Gennemsnitlig antal års skolegang, forældreparret	13,76	2,39	13,67	2,44	0,09	0,08
Gennemsnitlig antal års skolegang over gennemsnit, forældreparret	0,50	0,50	0,49	0,50	0,01	0,36

Anm.: Antal observationer varierer på de enkelte karakteristika mellem 2.003 og 2.246 for de elever, der deltager i forsøgsprøverne, og 54.110 og 61.353 for de elever, der ikke deltager i forsøgsprøverne. Forskellen mellem de to grupper og p-værdierne er beregnet ved brug af en t-test. De fremhævede tal indikerer, at forskellen mellem de to grupper er statistisk signifikant på minimum 5-%'s niveau.

Kilde: VIVEs analyser på baggrund af data fra Danmarks Statistik og STIL.

Tabellen viser, at de elever, som deltager i forsøget, på en række karakteristika er statistisk signifikant forskellige fra de elever, som ikke deltager i forsøget. Der er dog tale om små forskelle. Et eksempel herpå er den gennemsnitlige indkomst blandt de deltagende elevers forældre, som er en smule højere end blandt forældrene til de elever, som ikke deltager. Denne forskel er drevet af mødrene. Der er ingen signifikant forskel mellem de deltagende elevers fædres indkomst og fædrenes indkomst blandt de elever, som ikke deltager. Ser vi på de deltagende elever selv, er deres karaktergennemsnit i 8. klasse 6,76, mens det er 6,39 blandt de elever, som ikke deltager i forsøget. De deltagende elever har altså et karaktergennemsnit, som er 0,37 karakterpoint højere. Andelen af piger blandt de deltagende elever er 51 %, mens den

er 49 % blandt de elever, som ikke deltager. Det er en lille forskel på 2 procentpoint, der dog er statistisk signifikant. Ser vi på elevernes etnicitet, er andelen med dansk oprindelse den samme i de to grupper.

Selvom der er statistisk signifikante forskelle mellem de elever, som deltager i prøverne, og de elever, der ikke gør, er forskellene så små, at man kan argumentere for, at forskellene er uden betydning, og at de elever, der deltager i forsøget, er repræsentative for alle elever i 9. klasse.

En anden måde at tjekke, om eleverne, som deltager i forsøget, er repræsentative for elever i 9. klasse, er ved at definere en række elevgrupper og undersøge, om disse er at finde blandt de deltagende elever. Tabel 8.2 viser, hvor stor en andel de forskellige elevgrupper udgør af de elever, som deltager i forsøgsprøverne, og af de elever i 9. klasse, som ikke deltager i forsøgsprøverne.

Tabel 8.2 Repræsentation af forskellige elevgrupper

	Elever, som deltager i forsøget		Elever, som ikke deltager i forsøget		Forskel	P-værdi
Karakteristika	**Gennemsnit**	**Standard-afvigelse**	**Gennemsnit**	**Standard-afvigelse**		
Drenge med ikke-dansk oprindelse	0,06	0,23	0,06	0,24	0,00	0,40
Piger med ikke-dansk oprindelse	0,07	0,25	0,06	0,23	0,01	0,06
Elever, hvis forældres indkomst ligger i top 10 %	0,13	0,33	0,10	0,30	0,03	**0,00**
Elever, hvis forældres indkomst ligger i bund 10 %	0,11	0,31	0,10	0,30	0,01	0,25
Elever, hvis forældre har 17 eller flere års uddannelse (svarer til en kandidatuddannelse eller højere)	0,10	0,30	0,10	0,30	0,01	0,37
Elever, hvis forældre har 12 eller færre års uddannelse (svarer til højst en ungdomsuddannelse)	0,20	0,40	0,21	0,41	-0,01	0,18
Elevens karakterer i top 10 % i 8. klasse	0,12	0,33	0,10	0,30	0,02	**0,00**
Elevens karakterer i bund 10 % i 8. klasse	0,06	0,24	0,10	0,30	-0,04	**0,00**

Anm.: Antal observationer varierer på de enkelte karakteristika mellem 2.187 og 2.246 for de elever, der deltager i forsøgsprøverne, og 59.430 og 61.353 for de elever, der ikke deltager i forsøgsprøverne. Forskellen mellem de to grupper og p-værdierne er beregnet ved brug af en t-test. De fremhævede tal indikerer, at forskellen mellem de to grupper er statistisk signifikant på 5-%'s niveau.

Kilde: VIVEs analyser på baggrund af data fra Danmarks Statistik og STIL.

Vi finder, at alle de viste elevgrupper er repræsenteret både blandt de elever, som deltager i forsøget, og blandt dem, som ikke gør. Tabellen viser dog, at der er en smule forskel på, hvor stor en andel de enkelte elevgrupper udgør blandt hhv. de deltagende elever og de elever, som ikke deltager. Blandt de elever, som deltager i forsøget, er der en højere andel elever med relativt rige forældre, en højere andel med høje karakterer og en lavere andel elever med lave karakterer end blandt de elever, som ikke deltager i forsøget.

8.1.1 Skoler i hele landet er repræsenteret

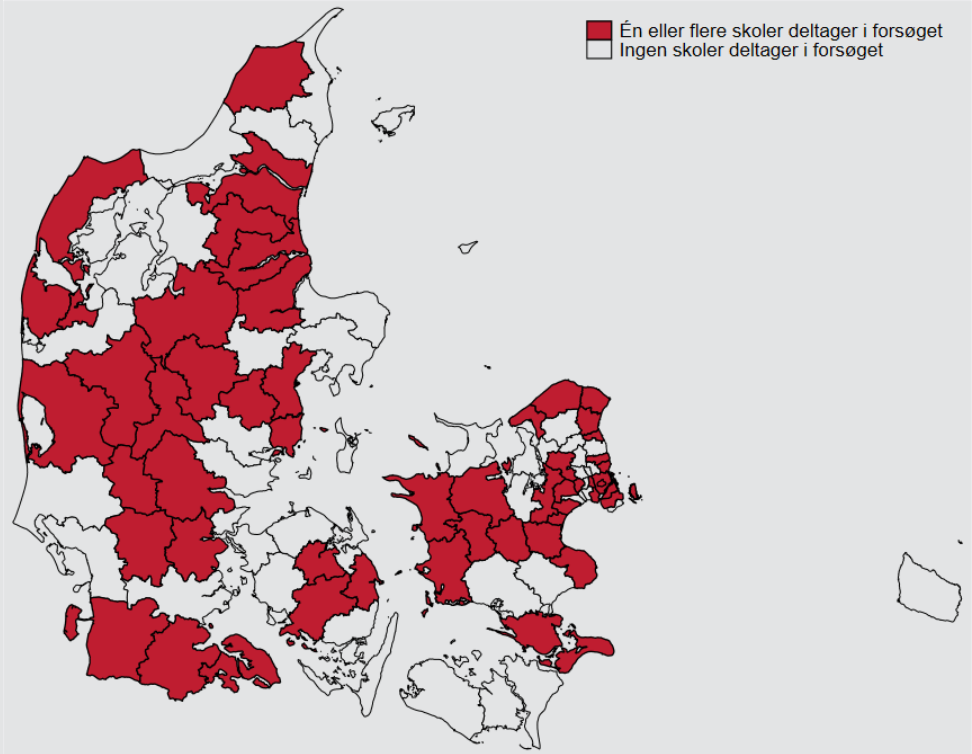
I alt deltager elever på 83 skoler i Danmark i forsøget[11]. Af disse er 62 folkeskoler (svarende til 75 %), 18 er fri- eller privatskoler (svarende til 22 %) og 3 er efterskoler (svarende til 3 %).

Ligesom det er vigtigt for generaliserbarheden af forsøgets resultater, at eleverne på en række karakteristika ligner den generelle population af elever i 9. klasse, og at alle typer elever er repræsenteret blandt de deltagende elever, er det ligeledes vigtigt, at skoler i forskellige dele af Danmark er repræsenteret.

Figur 8.1 viser de 53 kommuner, hvor der er minimum én deltagende skole i forsøget (markeret med rød). Fra figuren er det tydeligt, at skoler i alle dele af landet er repræsenteret.

[11] 83 skoler er de skoler, hvor der er elever, som har besvaret både del A og del B, og som derfor indgår i analyserne. Derudover er der to skoler, hvor eleverne ikke har besvaret både del A og del B. Disse er ikke medtaget i analyserne.

Figur 8.1 Kommuner med minimum én deltagende skole



Anm.: Figuren indeholder kun de skoler, hvor der er minimum én elev, der har besvaret både prøvens del A og del B.

Kilde: VIVEs analyser på baggrund af data fra STIL.

8.2 Prøvernes struktur og sværhedsgrad

I dette afsnit beskriver vi opbygningen af de tre prøver i hhv. biologi, fysik/kemi og geografi og analyserer deres sværhedsgrad. Vi ser først på den samlede sværhedsgrad af prøverne og dernæst på sværhedsgraden af de enkelte spørgsmål.

8.2.1 Forsøgsprøvernes opbygning

Forsøgsprøverne i biologi, geografi og fysik/kemi består hver især af en del A og en del B. Tabel 8.3 viser prøvernes opbygning.

Tabel 8.3 Prøvernes opbygning

	Biologi		Fysik/kemi		Geografi	
	Del A	**Del B**	**Del A**	**Del B**	**Del A**	**Del B**
Antal overordnede opgaver	19	3	22	3	21	3
Antal items	54	18	56	18	58	17
Antal mulige point	54	37	56	42	58	37

Anm.: Biologibiologiprøvens del A består egentlig af 20 overordnede opgaver og 57 items, men der opstod en fejl i prøveafviklingen, der betød, at BUVM besluttede at tage opgaven ud af opgavesættet. Derfor er der ingen elever, der har fået point i denne opgave. I delprøve A kan eleverne få op til 1 point for hvert rigtigt item. I delprøve B kan eleverne få mellem 1 og 7 point for et rigtigt item.

Kilde: VIVEs egen opstilling på baggrund af data fra STIL.

Tabel 8.3 viser, at prøverne i de tre fag ikke er helt identiske, hvad angår antal overordnede opgaver, antal underopgaver – eller items, som vi kalder dem – og det mulige antal point, eleverne kan opnå i prøverne. Det samlede antal items i del A varierer fra 54 til 58 og fra 17 til 18 i del B. Antal mulige point spænder fra 54 til 58 i prøvernes del A og fra 37 til 42 i prøvernes del B.

8.2.2 Pointfordelingen på tværs af delprøverne

Det gennemsnitlige antal point, som eleverne opnår i en prøve, giver et mål for prøvernes sværhedsgrad. Tabel 8.4 viser elevernes gennemsnitlige score i hvert fag og hver del.

Tabel 8.4 Gennemsnitlig score i hver delprøve i hvert fag

	Del A			Del B		
	Gennemsnit	Standard-afvigelse	Procentvis andel point ud af mulige point	Gennemsnit	Standard-afvigelse	Procentvis andel point ud af mulige point
Biologi	39,4	5,5	73 %	22,7	6,1	61 %
Fysik/kemi	36,8	7,4	66 %	21,4	8,9	51 %
Geografi	40,3	7,1	69 %	21,4	5,7	58 %

Anm.: Gennemsnittet i del B er fundet på baggrund af censor 1's vurdering.

Kilde: VIVEs analyser på baggrund af data fra STIL.

Tabel 8.4 viser, at elevernes gennemsnitlige score i del A er højest i geografiprøven, som imidlertid også er den prøve, hvor det er muligt at opnå flest point. For bedre at kunne sammenligne den gennemsnitlige score mellem de tre prøver, som har hvert sit maksimale antal point, dividerer vi den gennemsnitlige score i hver prøve med den højest mulige score. På den måde finder vi andel point i gennemsnit ud af mulige point. Når vi gør det, er andelen af point ud af mulige point højest i biologi, hvor eleverne i gennemsnit har fået 73 % ud af mulige point. Dernæst kommer geografi, hvor eleverne i gennemsnit har fået 69 % ud af mulige point, og endeligt fysik/kemi, hvor eleverne i gennemsnit har fået 66 % ud af mulige point.

Ser vi på prøvens del B og laver samme øvelse, kan vi se, at denne del generelt har været sværere for eleverne end del A, idet andel point ud af mulige point er lavere i alle tre prøver. Det kunne tyde på, at der er relativt få svære opgaver i del A, hvilket vi undersøger nærmere nedenfor. Ligesom i del A har eleverne fået flest point ud af mulige point i biologi, hvor de i gennemsnit har fået 61 % af de mulige point. Den sværeste prøve for eleverne har, også ligesom i del A, været fysik/kemi, hvor eleverne i gennemsnit har fået 51 % ud af det mulige antal point. I geografi har eleverne i gennemsnit fået 58 % af det samlede antal mulige point.

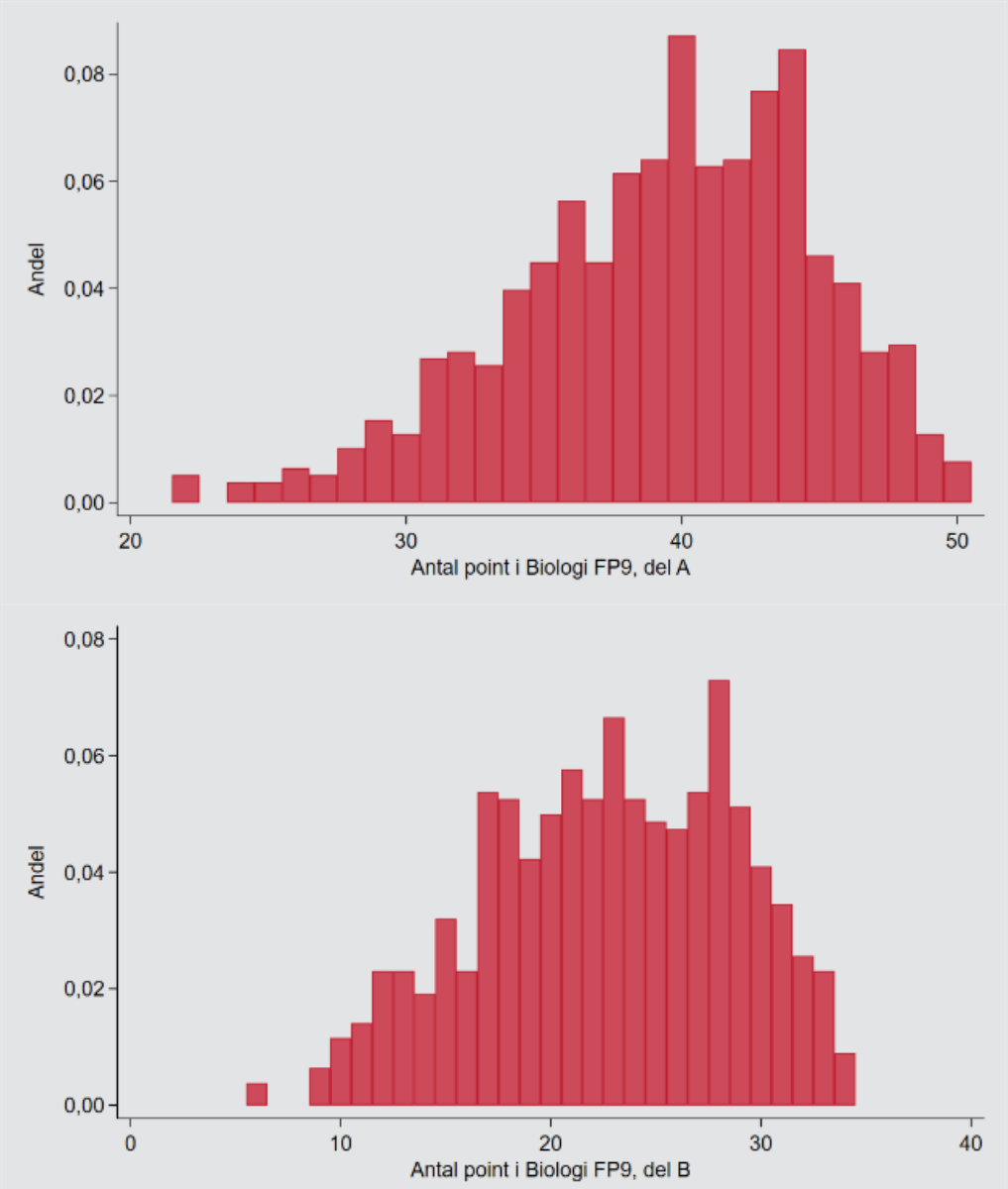
Vi ser nu nærmere på fordelingen af det samlede antal point i hver af de to dele i de tre fag.

Figur 8.2 viser fordelingen af det samlede antal point i prøvernes del A og B i biologi, Figur 8.3 i fysik/kemi og Figur 8.4 i geografi. Generelt viser fordelingerne det, vi også kan aflæse i Tabel 8.4. Nemlig at spredningen overvejende er større i del B end i del A, og at eleverne, som har taget biologiprøven, scorer et højere gennemsnitligt antal point end de elever, som har taget prøven i et af de andre fag. Dette kan vi se, fordi fordelingerne generelt er fladere og bredere for del B end del A, og fordi pointfordelingen for del A i biologi er mere højreskæv end fordelingerne i del A i de andre to fag.

Figur 8.2 viser, at pointfordelingen i del A i høj grad følger en normalfordeling, hvor der ligger flest elever omkring midten af fordelingen og færre i yderkanterne. Prøven fanger således elever på alle niveauer. Fordelingen for del B er derimod mere flad med flere elever ude i enderne af fordelingen. Dette indikerer, at denne prøve er bedre i stand til at identificere de meget dygtige og de mindre dygtige elever, end del A er.

Figur 8.2 Fordelingen af point i biologiprøvens del A og del B

Figuren viser fordelingen af antal point i biologiprøven. Øverste panel i figuren illustrerer point-fordelingen i del A, og nederste panel illustrerer pointfordelingen i del B.



Anm.: Hvis der er færre end tre elever, der har en bestemt score, er disse ikke medtaget i figuren på grund af diskretion.

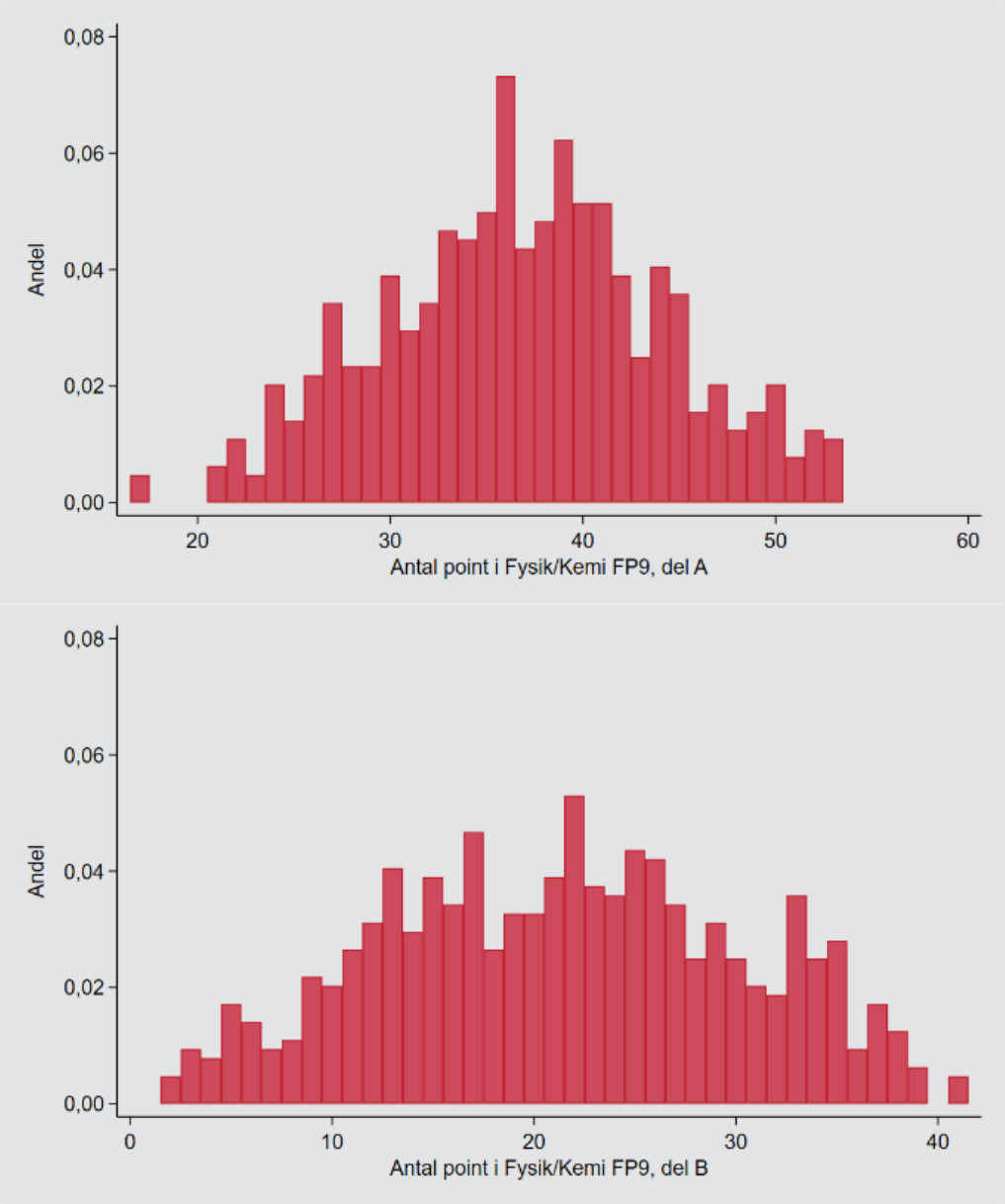
Kilde: Analyser af VIVE på baggrund af data fra STIL.

Figur 8.3 viser på samme måde for fysik/kemi, at pointfordelingen i del A er højere og smallere end fordelingen i del B, som er lavere og bredere. Her er

del B altså også i højere grad i stand til at identificere de meget dygtige og de mindre dygtige elever.

Figur 8.3 Fordelingen af point i fysik/kemi-prøvens del A og del B

Figuren viser fordelingen af antal point i fysik/kemi-prøven. Øverste panel i figuren illustrerer pointfordelingen i del A, og nederste panel illustrerer pointfordelingen i del B.



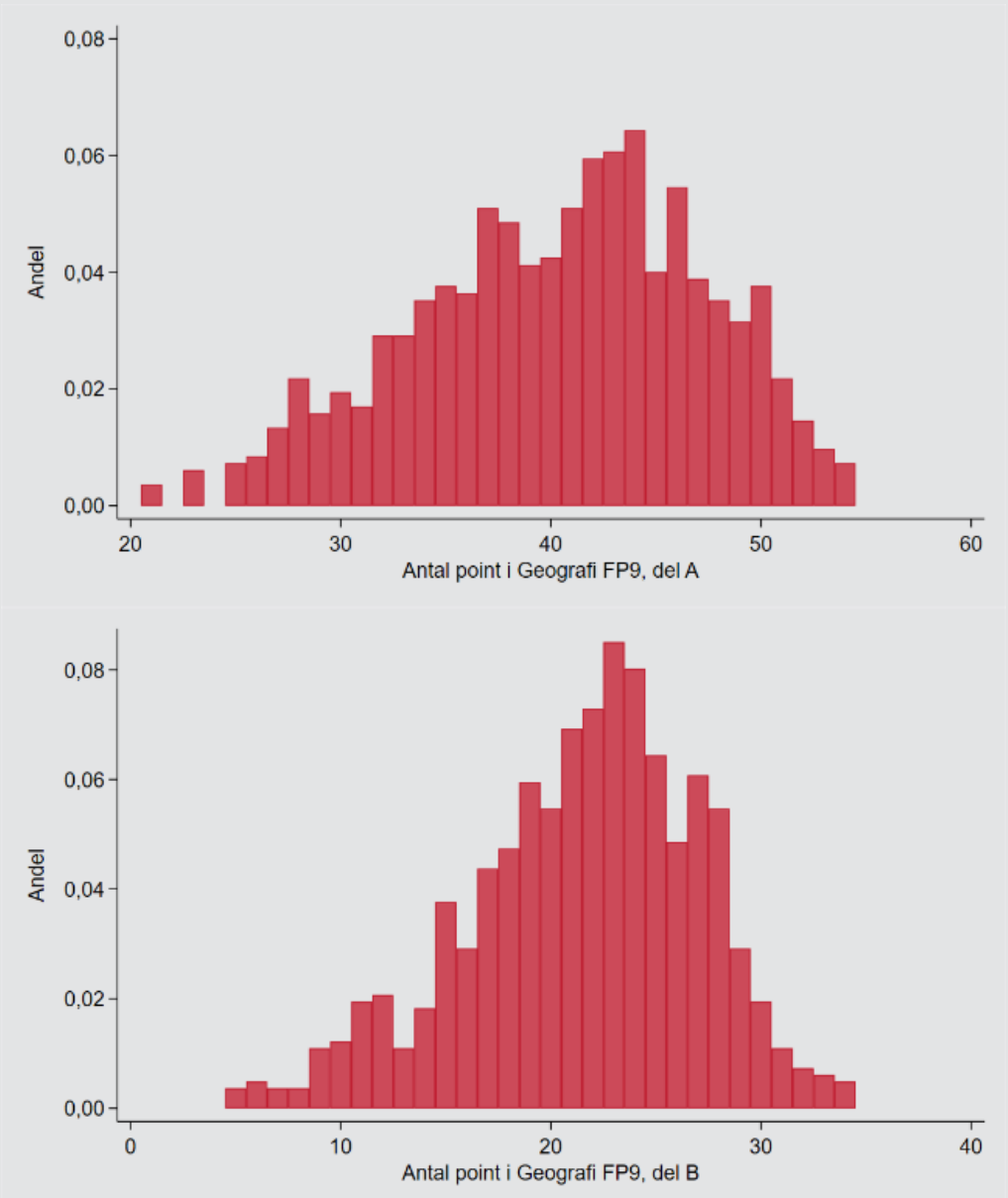
Anm.: Hvis der er færre end tre elever, der har en bestemt score, er disse ikke medtaget i figuren på grund af diskretion.

Kilde: Analyser af VIVE på baggrund af data fra STIL.

Figur 8.4 viser pointfordelingerne i del A og B i geografiprøven. Her tegner sig et lidt andet billede end for de andre to fag. Her er fordelingen højere og smallere i del B og lavere og bredere i del A.

Figur 8.4 Fordelingen af point i geografiprøvens del A og del B

Figuren viser fordelingen af antal point i geografiprøven. Øverste panel i figuren illustrerer pointfordelingen i del A, og nederste panel illustrerer pointfordelingen i del B.



Anm.: Hvis der er færre end tre elever, der har en bestemt score, er disse ikke medtaget i figuren på grund af diskretion.

Kilde: Analyser af VIVE på baggrund af data fra STIL.

8.2.3 Pointfordelingen på tværs af prøvespørgsmålene

I dette afsnit dykker vi længere ned i de enkelte prøver og undersøger, hvordan eleverne har klaret sig, når vi ser på de enkelte items.

Prøvernes del A er en multiple choice-prøve. Det vil sige, at eleverne enten kan svare rigtigt eller forkert og på hvert spørgsmål enten score 0 eller 1 point. Figur 8.5, Figur 8.6 og Figur 8.7 viser den gennemsnitlige score på hvert item i hver af de tre prøver. De enkelte items er sorteret, så items med den højeste gennemsnitlige score er øverst.

Det gør sig gældende for prøvens del A i alle tre fag, at der både er items, hvor næsten alle elever har svaret rigtigt, og items, hvor få elever har svaret rigtigt på det pågældende item. For alle tre prøver gælder det ligeledes, at items fordeler sig pænt, i forhold til hvor mange elever der svarer rigtigt på hvert item. Der er dog markant flere items, hvor den gennemsnitlige score ligger i toppen, dvs. mellem 0,8 og 1, end der er items, hvor den gennemsnitlige score ligger i bunden, dvs. mellem 0 og 0,2.

Figur 8.5 Gennemsnitlig score på hvert item, biologiprøven, del A

Figuren viser den gennemsnitlige score på hvert item i biologiprøvens del A.

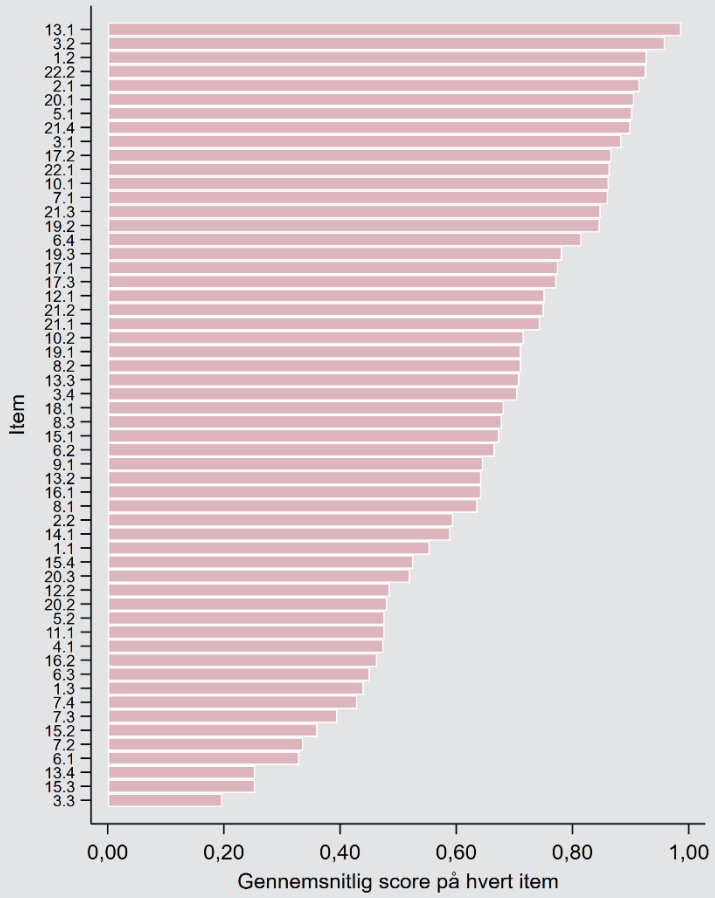
Kilde: Analyser af VIVE på baggrund af data fra STIL.

Figur 8.5 viser, at i biologiprøven er det item 7.1, som har den højeste gennemsnitlige score på 0,99, hvilket svarer til, at kun 8 elever ud af 780 ikke har svaret rigtigt på spørgsmålet. I den anden ende er det item 5.3, der har den laveste gennemsnitlige score på 0,10, hvilket svarer til, at 702 elever ikke har svaret rigtigt på spørgsmålet.

Figur 8.6 viser den gennemsnitlige score på hvert item i fysik/kemi-prøven. I denne prøve er det item 13.1, som har den højeste gennemsnitlige score på 0,99, hvilket svarer til, at det kun er 6 elever ud af 642, som ikke har svaret rigtigt på det item. Det item, som har den laveste gennemsnitlige score, og som færrest elever dermed har svaret rigtigt på, er item 3.3, som har en gennemsnitlig score på 0,20, hvilket svarer til, at 514 elever ikke har svaret rigtigt på det item.

Figur 8.6 Gennemsnitlig score på hvert item, fysik/kemi-prøven, del A

Figuren viser den gennemsnitlige score på hvert item i biologiprøvens del A.

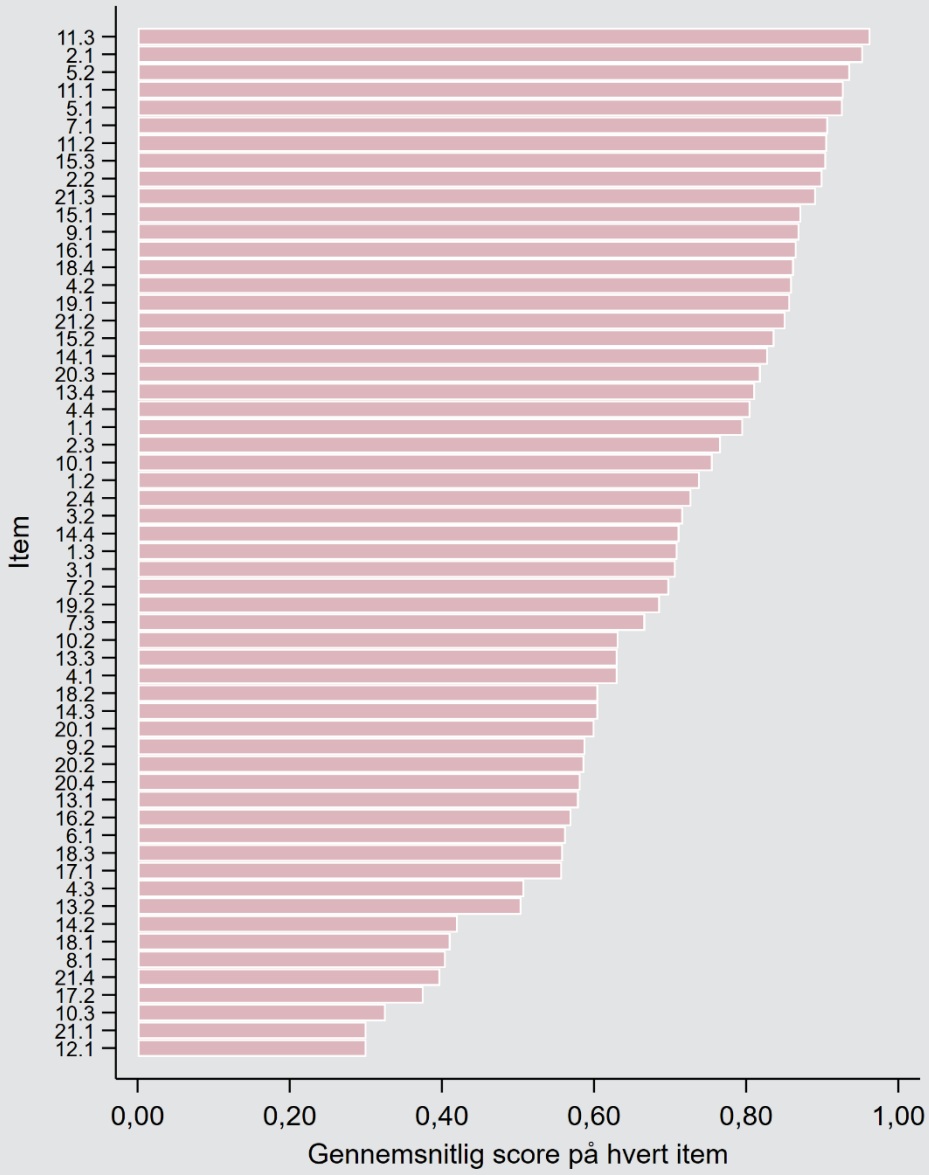


Kilde: Analyser af VIVE på baggrund af data fra STIL.

Figur 8.7 viser den gennemsnitlige score på hvert item i geografiprøven. Her er det item 11.3, som har den højeste gennemsnitlige score på 0,96, hvilket svarer til, at blot 33 elever ud af 824 ikke har svaret rigtigt på det item. Item 21.1 har den laveste gennemsnitlige score på 0,30, hvilket svarer til, at 577 elever ikke har svaret rigtigt på det item.

Figur 8.7 Gennemsnitlig score på hvert item, geografiprøven, del A

Figuren viser den gennemsnitlige score på hvert item i biologiprøvens del A.



Kilde: Analyser af VIVE på baggrund af data fra STIL.

I prøvernes del B er det forskelligt, hvor mange point eleverne kan opnå i det enkelte item. For at gøre de enkelte items sammenlignelige med hinanden og med del A har vi beregnet den gennemsnitlige score ud af den højest mulige score på hvert item. Figur 8.8 viser den gennemsnitlige score ud af den højest mulige score på hvert item i hver af de tre prøver. De items, hvor den gennemsnitlige score udgør den største andel af den højest mulige score, er længst til højre, hvor søjlerne er højest. De mørke søjler indikerer de opgaver i del B, hvor eleverne aktivt har skullet anvende simuleringer i opgavebesvarelsen, fx ved at følge instruktioner i at anvende simuleringen og derefter aflæse de korrekte oplysninger fra simuleringen.

Figur 8.8 Gennemsnitlig score på hvert item, del B

Figuren viser den gennemsnitlige score ud af den højest mulige på hvert item i prøvernes del B. Øverste panel i figuren illustrerer biologi, midterste panel illustrerer fysik/kemi, og nederste panel illustrerer geografi.

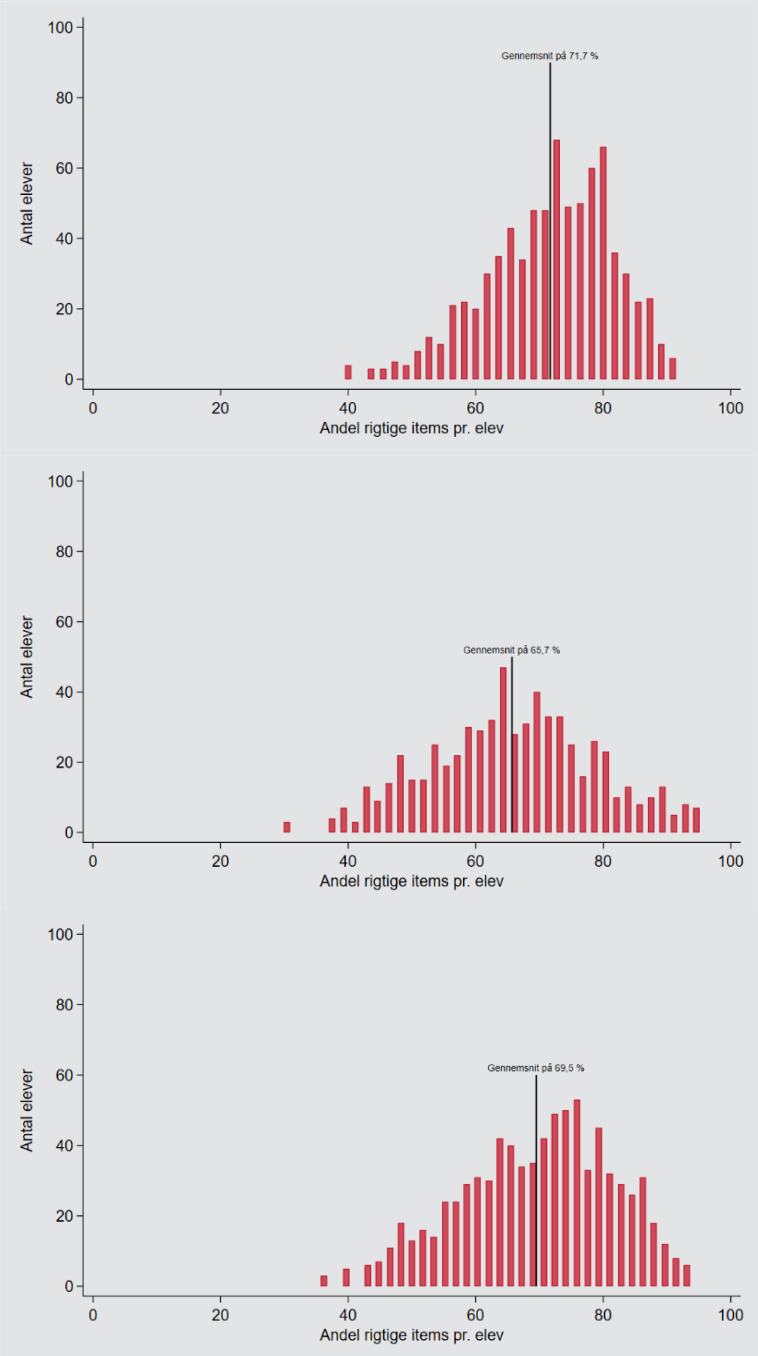
Kilde: Analyser af VIVE på baggrund af data fra STIL.

Ligesom for prøvernes del A er der både items, hvor den gennemsnitlige score er tæt på den højest mulige score, og items, hvor den gennemsnitlige score udgør en lille del af den mulige score på det item. De items, som ligger længst til højre, og hvor den gennemsnitlige score ud af den mulige score er højest, er generelt de første to items inden for en given opgave, og som for alle tre fag er en eller flere items i den af de tre hovedopgaver, hvor der anvendes simulering. I biologi og fysik/kemi er det opgave 1, mens det i geografi er opgave 3. Hver af hovedopgaverne, der handler om simulering, indeholder underspørgsmål eller items, der ikke hænger direkte sammen med simuleringen, mens eleverne i andre underspørgsmål har skullet anvende simuleringen direkte i besvarelsen af spørgsmålet. Det er disse, vi omtaler som simuleringsopgaver. I biologi er det item 1.1, 1.2, 1.3 og 1.4, som eleverne har nemmest ved, hvoraf de tre af dem er simuleringsopgaver. I fysik/kemi er det ligeledes den første simuleringsopgave, item 1.1, og i geografi de første to simuleringsopgaver, item 1.1 og 1.2, som eleverne har nemmest ved. De opgaver, som hører til den hovedopgave, der omhandler simulering, men hvor eleverne ikke direkte skal anvende simuleringen i besvarelsen af det enkelte item, har eleverne sværere ved end de opgaver, hvor de direkte anvender simuleringen.

Hvis vi undersøger, hvor stor en andel af alle items den enkelte elev svarer rigtigt på i hver delprøve, finder vi resultater, der understøtter dem præsenteret i afsnit 2.1, nemlig at biologiprøven er den prøve, eleverne har nemmest ved, og fysik/kemi-prøven den, som eleverne har sværest ved, og hvor der er mest spredning i fordelingen. Figur 8.9 viser fordelingen af den gennemsnitlige andel rigtige items for hver af de tre prøvers del A. Figuren viser, at den gennemsnitlige andel rigtige items er 71,7 % i biologi, 69,5 % i geografi og 65,7 % i fysik/kemi. Som vi også tidligere har set, er fordelingen for biologi højere og smallere end fordelingen for geografi og fysik/kemi. For alle tre fag gælder det, at meget få elever har mindre end 40 % rigtige items og relativt mange har 60 % eller flere rigtige items.

Figur 8.9 Fordelingen af gennemsnitlig andel rigtige items pr. elev, del A

Figuren viser fordelingen af den gennemsnitlige andel rigtige items for hver elev i prøvernes del A. Øverste panel i figuren illustrerer biologi, midterste panel illustrerer fysik/kemi, og nederste panel illustrerer geografi.

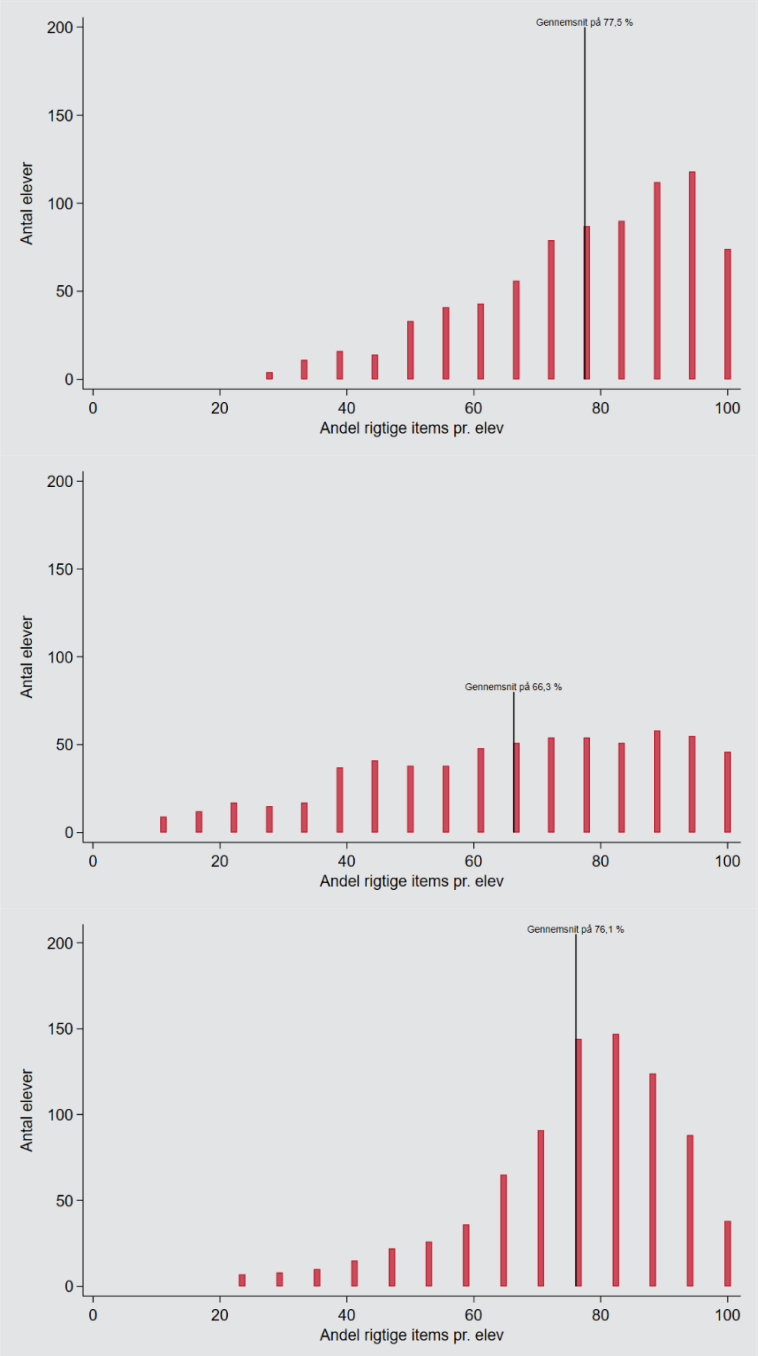


Kilde: Analyser af VIVE på baggrund af data fra STIL.

Ser vi på prøvernes del B og fordelingen af andel rigtige items pr. elev, tegner der sig meget det samme mønster som i prøvernes del A. Figur 8.10 viser fordelingen af det gennemsnitlige andel rigtige items for hver af de tre prøvers del B. Det gennemsnitlige andel rigtige items er højest i biologi, hvor eleverne i gennemsnit svarer rigtigt på 77,5 % af alle items. I geografi svarer eleverne i gennemsnit rigtigt på 76,1 % af alle items, og i fysik/kemi svarer eleverne i gennemsnit rigtigt på 66,3 % af alle items. Fordelingen for fysik/kemi er relativt bred og flad, hvilket indikerer, at der er omtrent lige mange elever, som har svaret rigtigt på mange og få items. Fordelingerne for biologi og geografi er begge mere højreskæve, hvilket indikerer, at der er flere elever, som har svaret rigtigt på en høj andel items end på en lav andel items.

Figur 8.10 Fordelingen af gennemsnitlig andel rigtige items pr. elev, del B

Figuren viser fordelingen af den gennemsnitlige andel rigtige items for hver elev i prøvernes del B. Øverste panel i figuren illustrerer biologi, midterste panel illustrerer fysik/kemi, og nederste panel illustrerer geografi.



Anm.: Analyserne af del B er baseret på censor 1's vurdering.

Kilde: Analyser af VIVE på baggrund af data fra STIL.

I prøvernes del B varierer det, hvor mange point eleverne kan score i hvert enkelt item. Vi ser derfor nærmere på fordelingen inden for hvert item i hver prøve.

Tabel 8.5 viser fordelingen af point på de enkelte items i biologiprøven.

Tabel 8.5 Fordeling af point på de enkelte items. Biologi, del B. Procent.

Item/point	0	1	2	3
1.1	4	96	-	-
1.2	4	1	2	93
1.3	7	14	79	-
1.4	1	3	44	52
1.5	14	38	49	-
1.6	17	59	24	-
1.7	22	49	29	-
2.1	1	52	47	-
2.2	24	27	49	-
2.3	26	47	27	-
2.4	29	23	23	25
2.5	43	57	-	-
2.6	20	47	33	-
3.1	33	67	-	-
3.2	40	39	21	-
3.3	30	27	25	18
3.4	31	33	37	-
3.5	58	34	8	-

Kilde: Analyser af VIVE på baggrund af data fra STIL.

Tabel 8.5 viser, at eleverne i biologiprøvens del B kan score op til 2 point på langt de fleste items, op til 3 point på en håndfuld items og op til 1 point på tre items. På nogle items scorer næsten alle eleverne maksimumpoint. Det gælder eksempelvis item 1.1, hvor 96 % af eleverne får 1 ud af 1 mulige point, item 1.2, hvor 93 % af eleverne får 3 ud af 3 mulige point, og item 1.3, hvor 79 % får 2 ud af 2 mulige point. Generelt er der en lille andel elever, som scorer 0 point i alle items i opgave 1. Der er til gengæld også items, hvor den største andel ikke svarer rigtigt. Det drejer sig om item 3.5, som er prøvens sidste item, hvor 58 % af eleverne scorer 0 point. Generelt er andelen, som får 0 point, højere

blandt alle items i opgave 3 end i de to resterende opgaver. Det kan enten betyde, at opgave 3 har været sværere, eller at eleverne løber tør for tid til den sidste del.

I fysik/kemi-prøvens del B kan eleverne i hovedparten af de enkelte items score op til 2 point. Og i fire opgaver op til enten 3, 4 eller 7 point. Tabel 8.6 viser fordelingen af point for hvert enkelt item.

Tabel 8.6 Fordeling af point på de enkelte items. Fysik/kemi, del B. Procent.

Item/ point	0	1	2	3	4	5	6	7
1.1	4	96	-	-	-	-	-	-
1.2	51	7	43	-	-	-	-	-
1.3	34	66	-	-	-	-	-	-
1.4	48	7	4	42	-	-	-	-
1.5	26	29	45	-	-	-	-	-
1.6	29	48	23	-	-	-	-	-
1.7	38	35	28	-	-	-	-	-
2.1	3	3	13		82[1]	-	-	-
2.2	10	18	21	20	16	9	4	3
2.3	58	12	30	-	-	-	-	-
2.4	46	30	24	-	-	-	-	-
2.5	11	15	22	25	27	-	-	-
3.1	24	34	42	-	-	-	-	-
3.2	29	71	-	-	-	-	-	-
3.3	45	37	18	-	-	-	-	-
3.4	48	38	14	-	-	-	-	-
3.5	48	52	-	-	-	-	-	-
3.6	55	37	8	-	-	-	-	-

Note: [1]Andelen med 3 og 4 point i item 2.1 er slået sammen på grund af diskretion.

Kilde: Analyser af VIVE på baggrund af data fra STIL.

Generelt er der mere spredning i fordelingen af point inden for de enkelte items i denne prøve end i biologiprøven. I fysik/kemi-prøven er der færre items, hvor meget få elever har scoret 0 point. Det gælder kun item 1.1, som 4 % ikke har svaret rigtigt på, og item 2.1, som 3 % ikke har svaret rigtigt på. På samme vis er der også kun to items, item 2.1 og 1.2, hvor næsten alle ele-

ver har scoret det maksimale antal point. Der er ét item, item 2.2, hvor eleverne har kunnet score op til 7 point. Det er en meget lille andel, som scorer 6 eller 7 point (hhv. 4 og 3 %).

Tabel 8.7 viser endeligt fordelingen af point på de enkelte items i geografiprøvens del B.

Tabel 8.7 Fordeling af point på de enkelte items. geografi, del B. %

Item/point	0	1	2	3	4
1.1	20	81	-	-	-
1.2	10	29	37	25	-
1.3	65	35	-	-	-
1.4	23	77	-	-	-
1.5	8	14	42	25	12
2.1	11	89	-	-	-
2.2	1	2	21	75	-
2.3	47	37	16	-	-
2.4	57	26	17	-	-
2.5	10	35	44	11	-
3.1	5	95	-	-	-
3.2	2	8	90	-	-
3.3	13	14	28	17	28
3.4	20	37	43	-	-
3.5	54	46	-	-	-
3.6	47	36	17	-	-
3.7	15	38	29	11	7

Kilde: Analyser af VIVE på baggrund af data fra STIL.

Eleverne har i denne prøve kunnet score mellem 1 og 4 point afhængigt af item. Tabellen viser, at der er tre items, som meget få elever ikke har svaret rigtigt på. Det drejer sig om item 2.2 og item 3.2, hvor blot 1 og 2 % ikke opnår minimum 1 point, og item 3.1, hvor 5 % ikke opnår minimum ét point. På samme måde er der også en lille håndfuld items, hvor en stor andel elever scorer det maksimale antal point. På de tre items, hvor eleverne har kunnet score op til 4 point, følger pointgivningen en normalfordeling med den højeste andel i midten og en mindre andel i enderne.

Generelt viser fordelingerne for hvert item i de tre prøver, at der både er items, som mange svarer rigtigt på og scorer topkarakter på, items, som en lavere andel scorer maksimumpoint på, og items, som er tættere på normalfordelt.

8.3 Prøvernes grundlæggende egenskaber

I dette afsnit præsenterer vi uddybende detaljer fra analyserne af prøvernes grundlæggende egenskaber. Vi viser først udvalgte resultater fra estimering af Rasch-modellerne, og derefter præsenterer vi detaljerede resultater for de figurer, der er vist i afsnit 2.2.

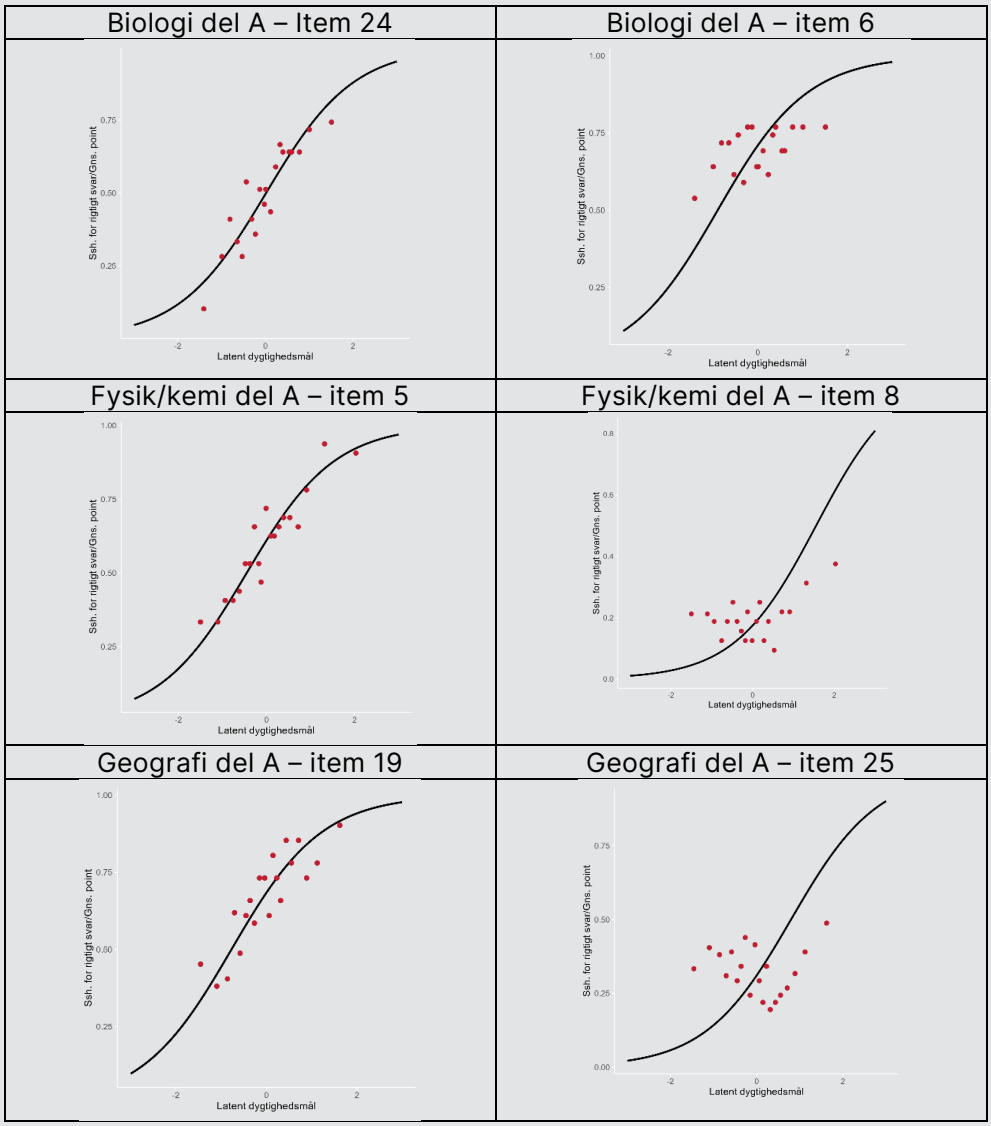
8.3.1 Estimering af Rasch-modellen

Vi har estimeret seks Rasch-modeller ved brug af pakken TAM i programmet R version 4.2.2. De tre del A-prøver i biologi, fysik/kemi og geografi er estimeret som dikotome Rasch-modeller, og de tre del B-prøver i biologi, fysik/kemi og geografi er estimeret som PCM Rasch-modeller.

Rasch-modellen er tæt forbundet med de tidligere statistiske analyser, hvor gennemsnitligt antal point for et spørgsmål er et simpelt mål for spørgsmålets sværhedsgrad, og antal point, som en elev får, er et mål for elevens færdigheder. I en Rasch-model modelleres begge dele dog simultant, således at alle opgavers sværhedsgrader vurderes simultant med alle elevers færdigheder. Det er derfor en fundamental egenskab (og antagelse) ved en Rasch-model, at sandsynligheden for, at en elev svarer rigtigt på et spørgsmål, kun afhænger af de underliggende færdigheder og spørgsmålets sværhedsgrad. Da både sværhedsgrad og færdigheder måles i 'logit-enheder', kan de direkte sammenlignes, og hvis en elev forsøger at svare på et spørgsmål med samme sværhedsgrad som elevens færdigheder, vil sandsynligheden for, at eleven svarer rigtigt, være 50 %.

Figur 8.11 Rasch-modellens forudsagte sandsynlighed for korrekte elevbesvarelser og faktiske elevbesvarelser for seks udvalgte items

Disse figurer betegnes som Item-Characteristic Curves (ICC).



Anm.: De sorte linjer angiver modellens forudsigelser, og de røde prikker angiver elevernes faktiske besvarelser. Hver røde prik repræsenterer fem % af data.

Kilde: VIVEs analyser af data fra STUK og data fra Danmarks Statistik.

Rasch-analysen giver et estimat af hvert items sværhedsgrad. Vi kan illustrere denne sværhedsgrad ved en figur, der viser sandsynligheden for at svare rigtigt på det givne item på y-aksen som funktion af elevernes dygtighed på x-

aksen. Vi kan dernæst sammenholde figuren med elevernes besvarelser givet elevernes estimerede dygtigheder. De fleste figurer viser en pæn sammenhæng mellem modellens forudsigelse og de faktiske data. Som illustration viser vi i Figur 8.11 et eksempel på en god overensstemmelse (det venstre panel) og et eksempel på en dårlig overensstemmelse i hvert af de tre fag.[12]

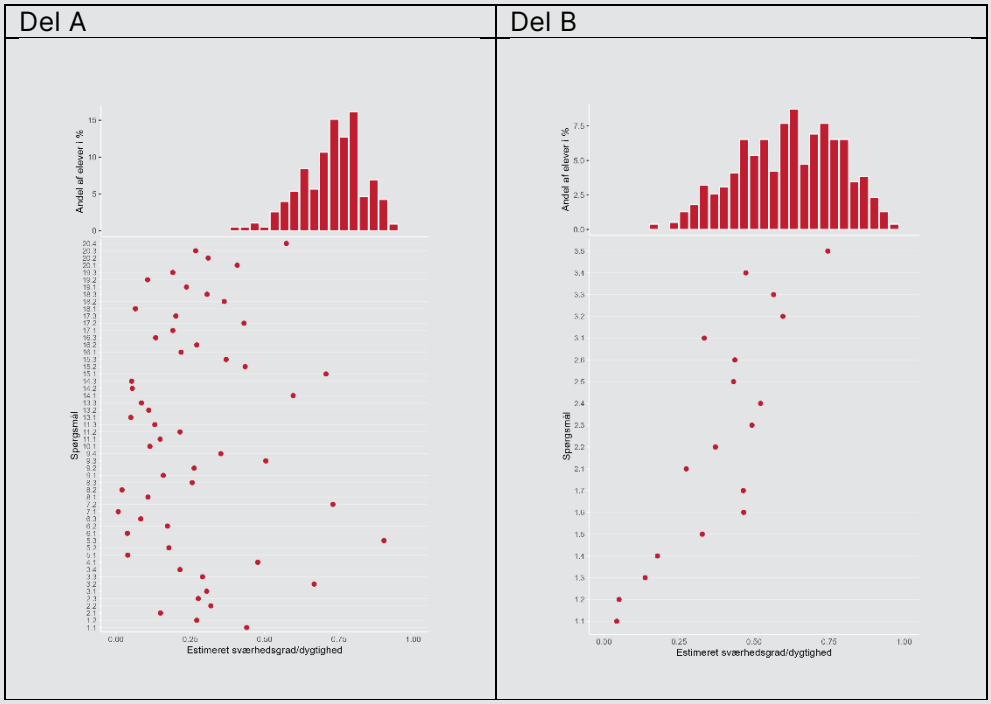
Sammenhængen mellem de estimerede sværhedsgrader og dygtigheder kan også illustreres ved et såkaldt Wright Map, som viser både elevernes estimerede dygtighed (øverst) og hvert items estimerede sværhedsgrad (nederst). Både dygtighed og sværhedsgrad er angivet i såkaldte logit-enheder. Sammenligneligheden betyder, at hvis en elev har samme logit-værdi i dygtighed, som et spørgsmål har i sværhedsgrad, så er der jf. Rasch-modellen 50 % chance for, at eleven svarer rigtigt på et spørgsmål.

Figur 8.12 til Figur 8.14 viser Wright Maps for de tre fag. Ligesom analyserne i det foregående afsnit viser disse figurer, at del B typisk er lidt sværere end del A, idet item-sværhedsgraden ligger længere til højre for den del. Overordnet ser vi, at på tværs af de to dele er der items, der repræsenterer hele spektret af sværhedsgrader og dermed også afdækker hele variationen i elevdygtighed.

[12] Samlet set er der 221 figurer af denne type, og det er derfor ikke hensigtsmæssigt at afrapportere dem alle her. Vi har derfor udvalgt eksempler fra hver forsøgsprøve som illustration.

Figur 8.12 Estimater for elevdygtighed og item-sværhedsgrad i biologi

Disse figurer betegnes som Wright Maps.

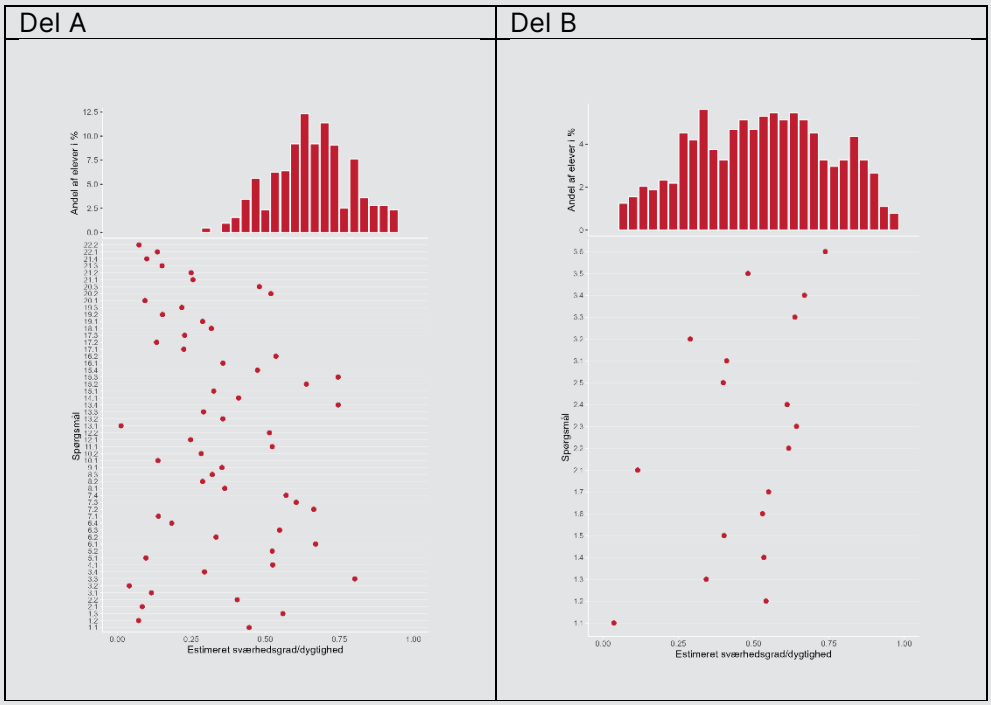


Anm.: Enhederne på x-aksen er logit.

Kilde: VIVEs analyser på data fra STUK og Danmarks Statistik.

Figur 8.13 Estimater for elevdygtighed og item-sværhedsgrad i fysik/kemi

Disse figurer betegnes som Wright Maps.

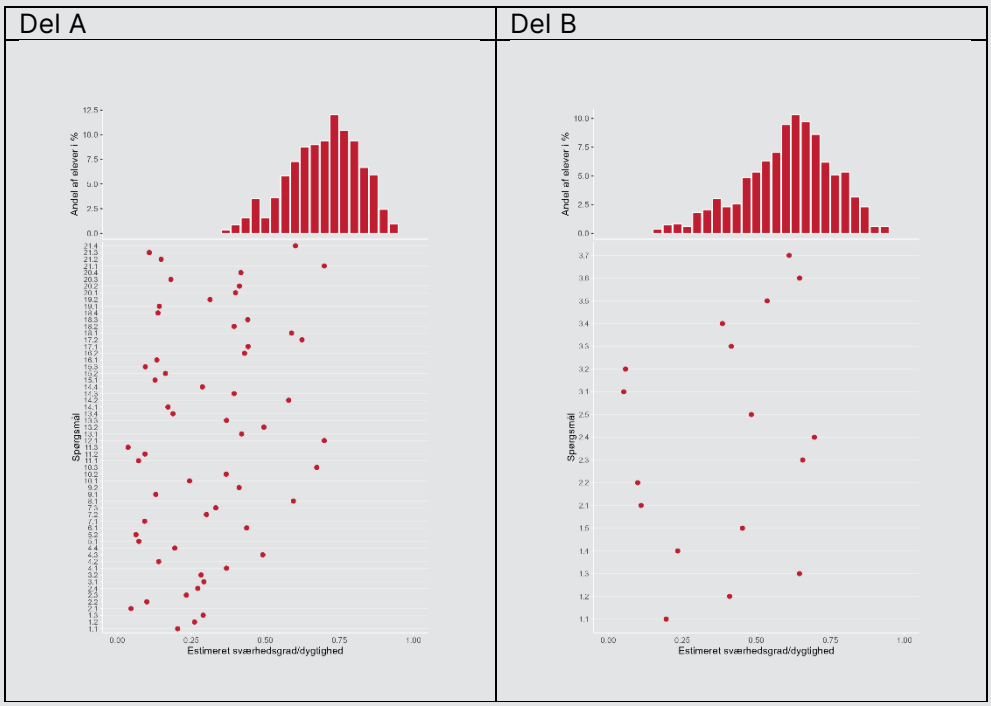


Anm.: Enhederne på x-aksen er logit.

Kilde: VIVEs analyser på data fra STUK og Danmarks Statistik.

Figur 8.14 Estimater for elevdygtighed og item-sværhedsgrad i geografi

Disse figurer betegnes som Wright Maps.



Anm.: Enhederne på x-aksen er logit.

Kilde: VIVEs analyser på data fra STUK og Danmarks Statistik.

Overordnet set peger ICC-figurerne og de tre Wright Maps på en rimelig overensstemmelse mellem data og Rasch-modellen. Der er dog flere items, der afviger en smule. Dette vil også være synligt i de følgende analyser, hvor vi afdækker egenskaber af den estimerede Rasch-model.

8.3.2 Item-stabilitet

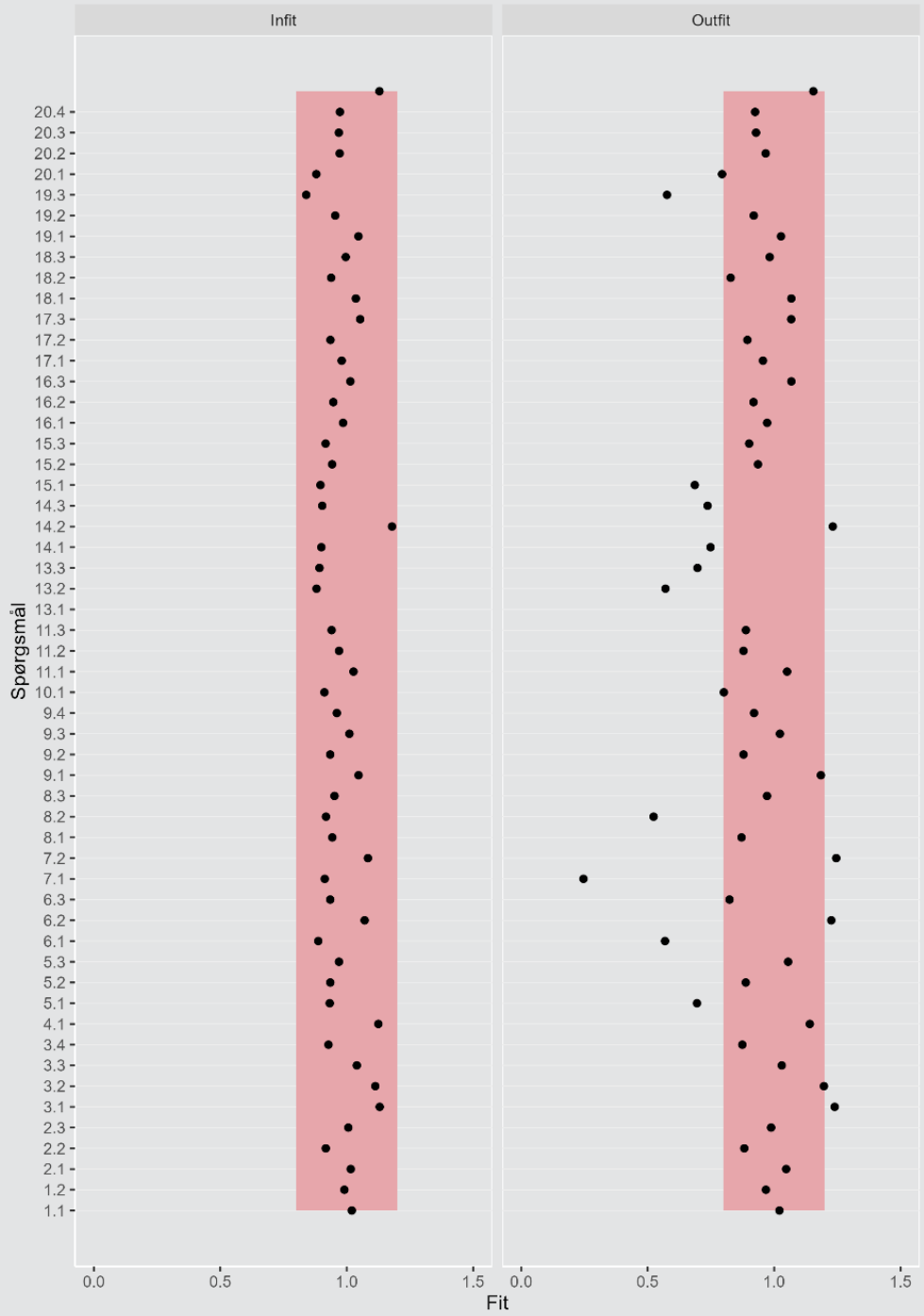
Infit og outfit er et mål for variationen i besvarelserne på et spørgsmål. Et lavt fit tyder på, at der er mindre variation end forventet, og et højt fit, at der er mere variation end forventet.

Figur 8.15 til Figur 8.20 viser infit og outfit på hvert item på tværs af alle tre fag og begge dele. Et for højt fit betyder, at der er lidt for meget variation i et spørgsmåls besvarelser, og et for lavt fit, at der er for lidt variation. Infit-målet vægter besvarelser, der ligger tættere på det enkelte items niveau højere, og det er derfor mindre følsomt over for outliers. For high-stakes-prøver som disse vil man almindeligvis kræve, at de ligger mellem 0,8 og 1,2 (Wright & Linacre, 1994).

Som figurerne viser, ligger langt de fleste items inden for normalområdet angivet som det røde felt. Som forventet er der flere items uden for normalområdet med outfit, som jo er mere følsomt overfor ekstreme værdier. Mens der overordnet set ikke er tegn på, at andelen af værdierne, der ligger uden for normalområdet, er højere i del B sammenlignet med del A, er det tydeligt, at der for enkelte items i del B er nogle meget store afvigelser.

Figur 8.15 Infit og outfit for del A i biologi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers.

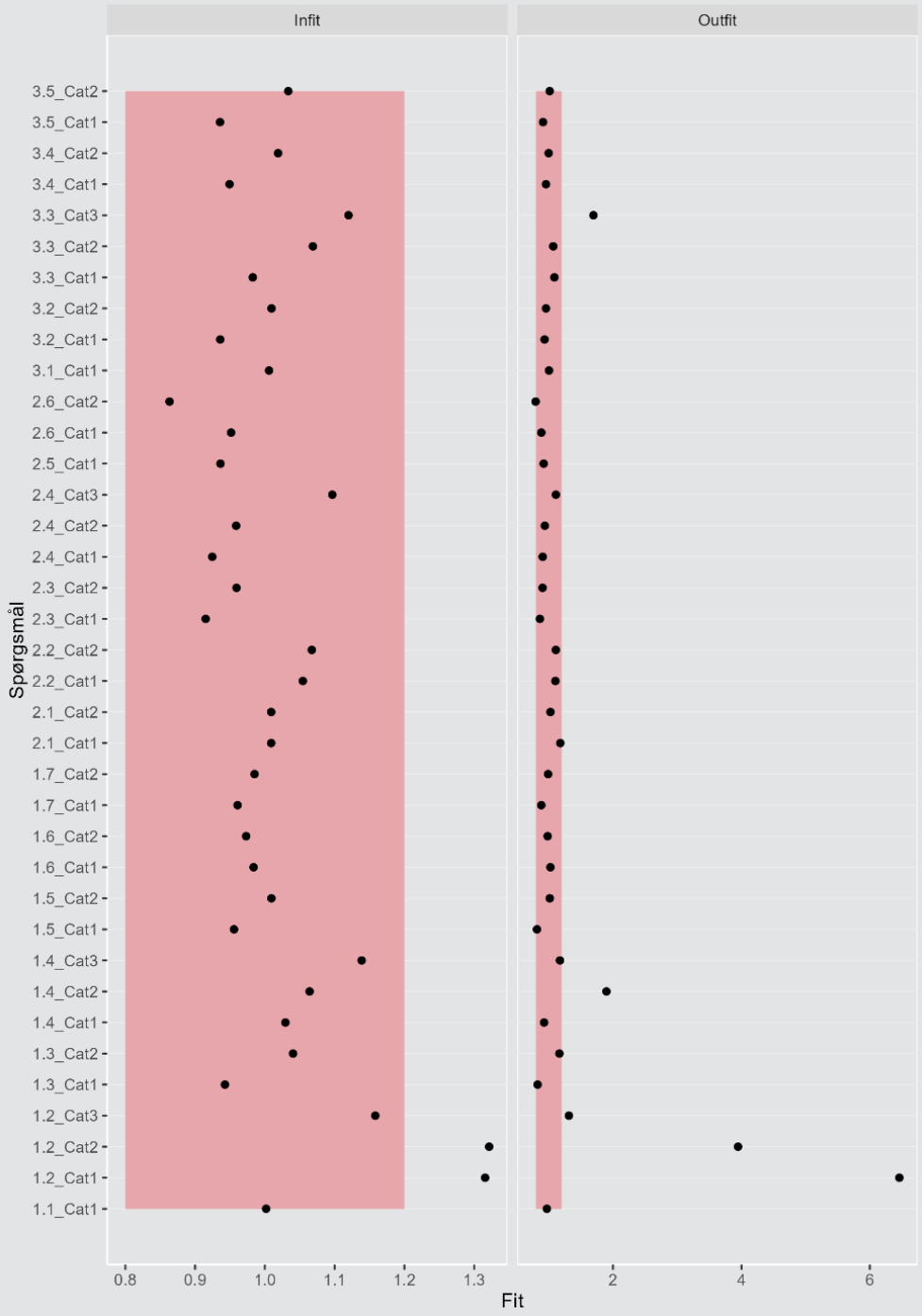


Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

Figur 8.16 Infit og outfit i del B i biologi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers.

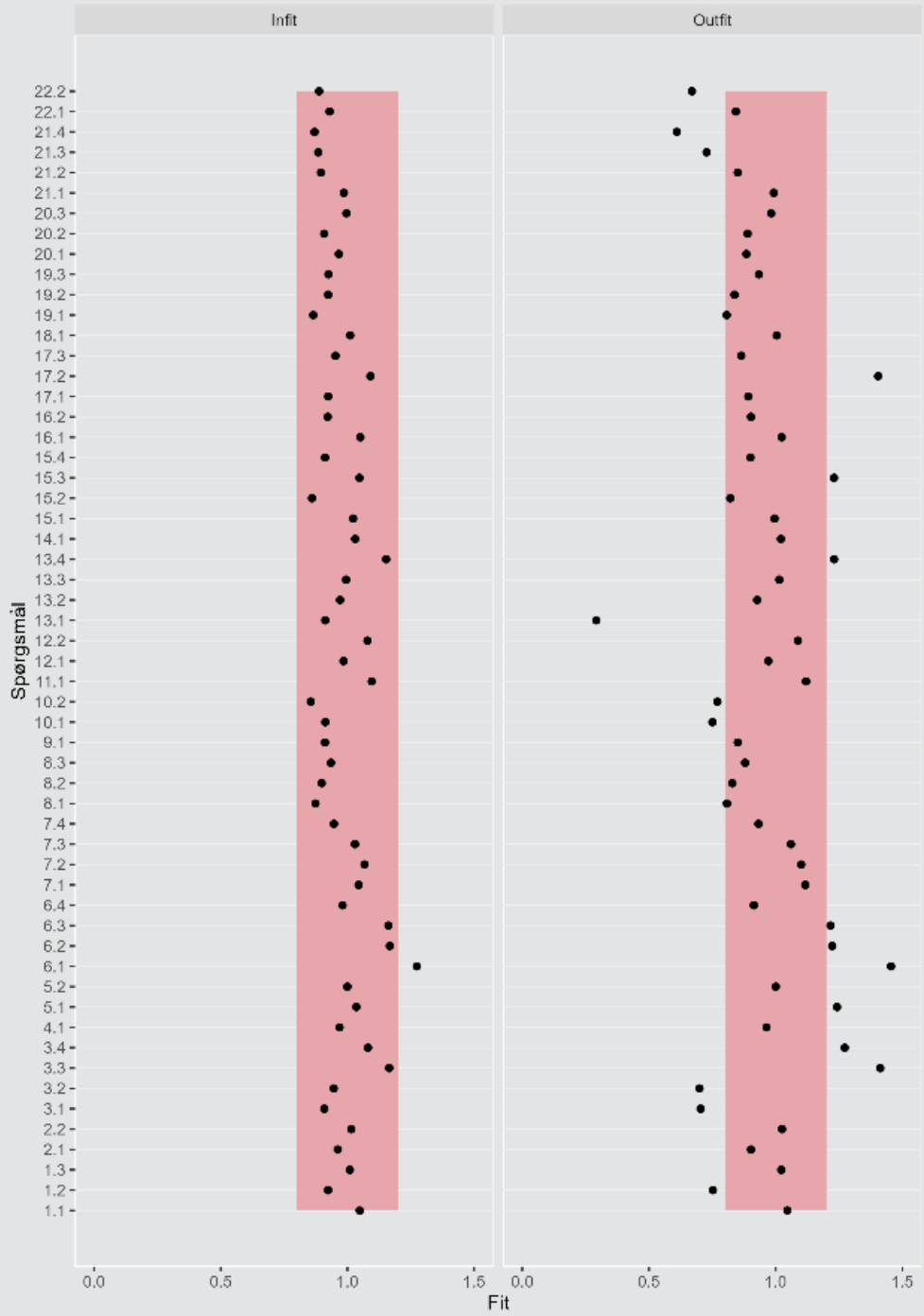


Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

Figur 8.17 Infit og outfit i del A i fysik/kemi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers.

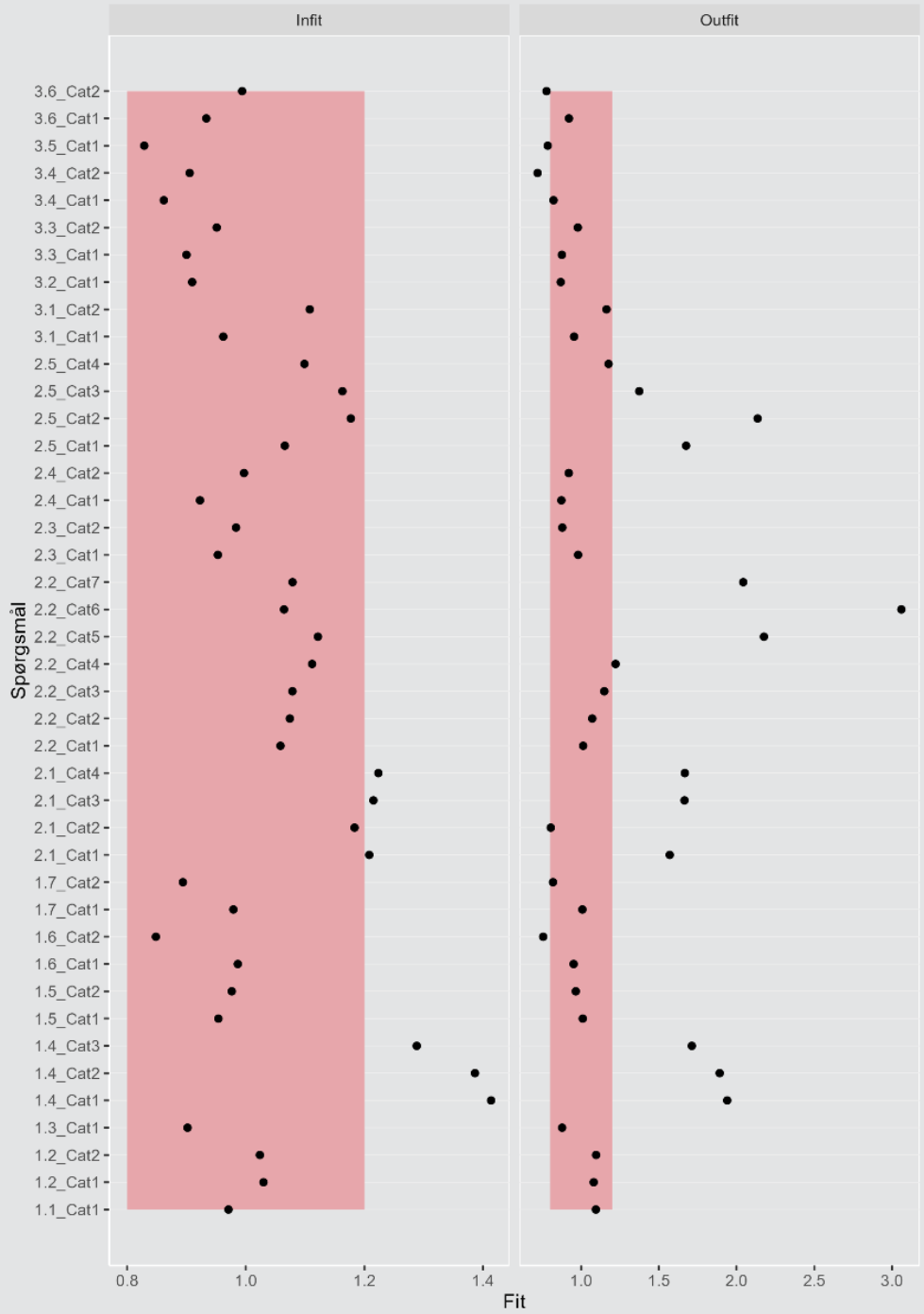


Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

Figur 8.18 Infit og outfit i del B i fysik/kemi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers.

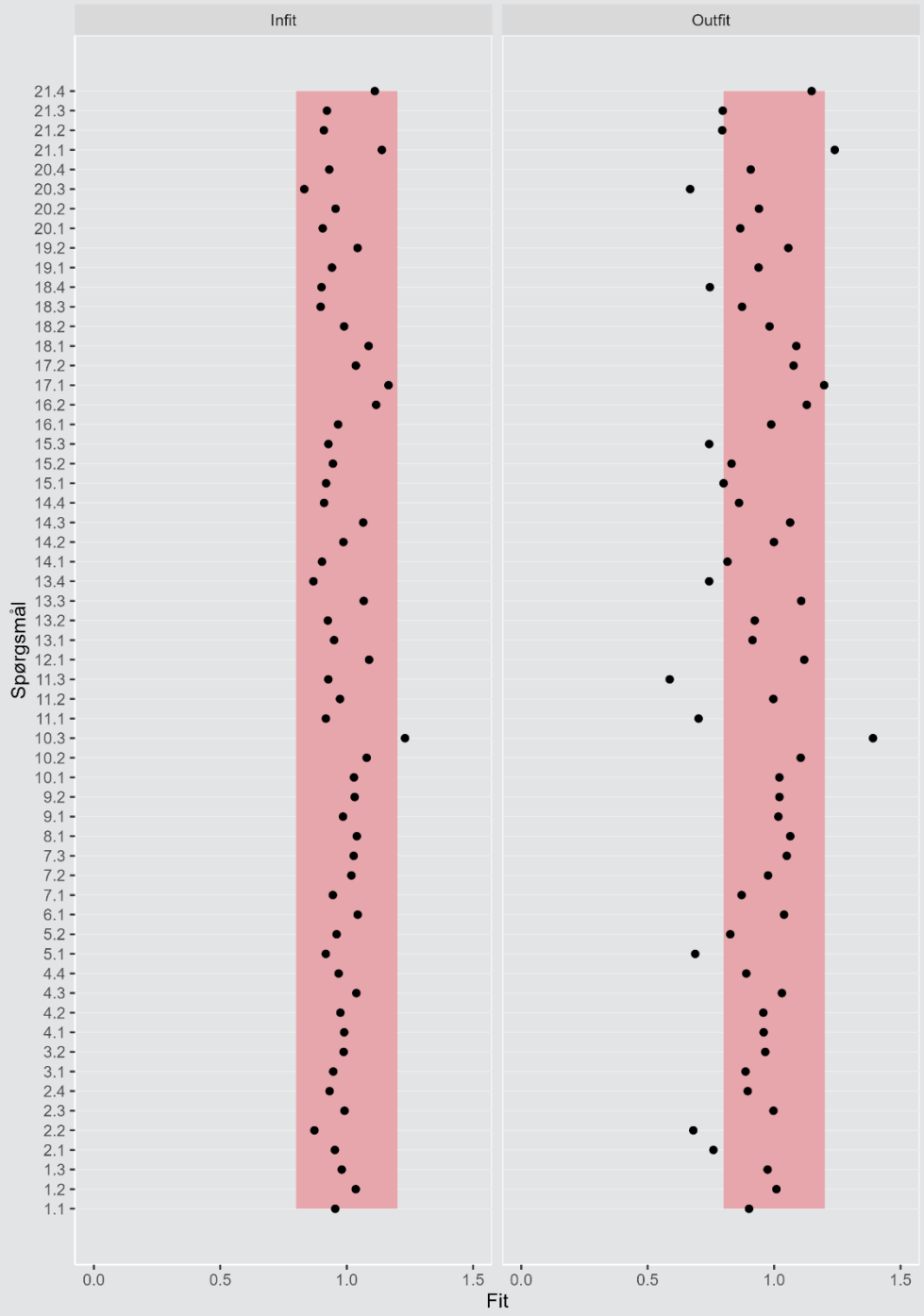


Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

Figur 8.19 Infit og outfit i del A i geografi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers

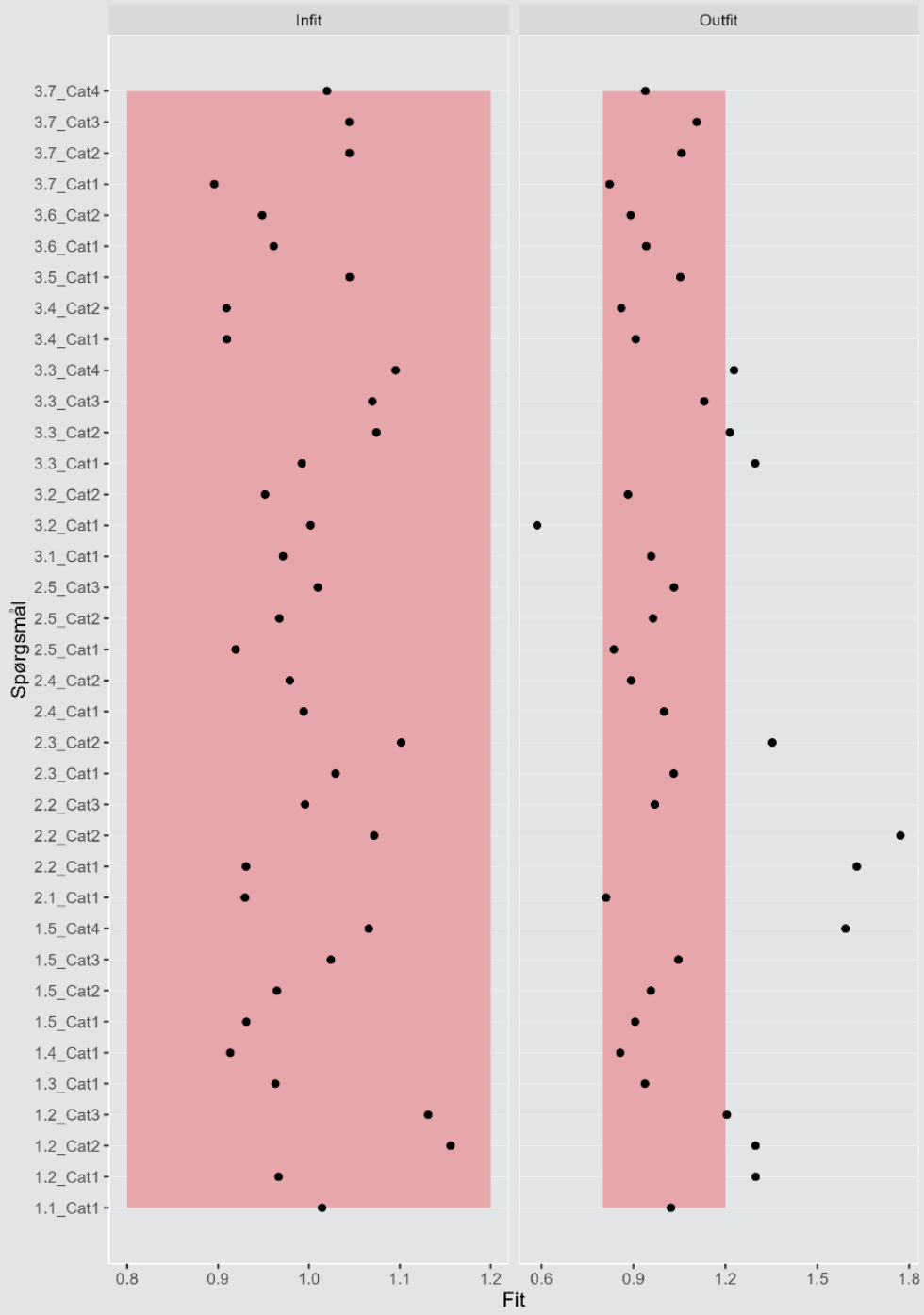


Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

Figur 8.20 Infit og outfit i del B i geografi

Infit måler items stabilitet ved at vægte besvarelser tæt på det items sværhedsgrad højest. Outfit vægter alle observationer ens og er derfor mere følsomt over for outliers.



Anm.: De røde områder angiver normalområdet på 0,8 til 1,2.

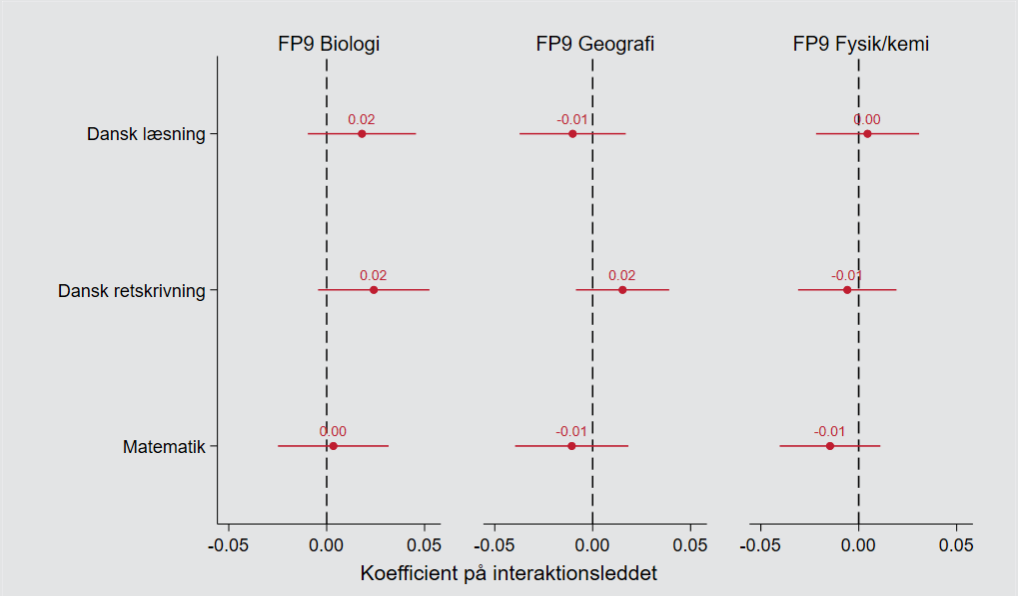
Kilde: Egne beregninger på data fra STUK og Danmarks Statistik.

8.4 Sammenhæng mellem prøveresultater og elevbaggrund

Hvor vi i Figur 2.4 viste sammenhængen mellem elevernes gennemsnitlige score i de øvrige prøver og forsøgsprøverne, så viser Figur 8.21 sammenhængen mellem resultaterne i de øvrige prøver og sandsynligheden for at være blandt de bedste 20 % i forsøgsprøverne. Heller ikke her ser vi tegn på, at danskfærdigheder har en særlig betydning for de nye delprøver.

Figur 8.21 Sammenhæng mellem elevernes resultater i øvrige prøver og sandsynligheden for at være blandt de bedste 20 % i prøvernes del A og del B

Figuren viser koefficienterne (de røde prikker) og 95-%'s konfidensintervallet (de røde linjer) for om sammenhængen med del B er højere end sammenhængen med del A. En positiv sammenhæng betyder, at eleverne scorer højere point i del A end i del B. Rører konfidensintervallet de stiplede lodrette linjer er sammenhængen *ikke* statistisk signifikant.



Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Vi kan også benytte Rasch-modellen til at undersøge, om prøven virker forskelligt på tværs af to elevgrupper. Det vil sige, at vi undersøger, om der er

Differential Item Functioning (DIF). Det gør vi ved at estimere modellen separat for hver gruppe og så sammenligne de estimerede sværhedsgrader. Tabel 8.8 viser alle items, hvor forskellen på tværs af grupperne er større end 0,5 logit-enheder. Det er forventeligt, at de estimerede sværhedsgrader er forskellige på tværs af grupper, da de forskellige grupper har forskellige forudsætninger for at svare rigtigt. Det er dog værd at bemærke, at benytter vi del A som benchmark, er andelen af items med forskelle i logit på over 0,5 generelt ikke højere i den ny prøvedel, del A.

Tabel 8.8 Items med forskel i logit-enheder på over 0,5 på tværs af de angivne grupper

Fag	Del	Dimension	Antal items	Antal items med DIF > 0,5	Items med DIF > 0,5
Biologi	A	Dansk	54	27	19.2(2.61), 18.1(2.02), 6.3(1.76), 2.2(1.59), 13.2(1.46), 19.3(1.41), 1.2(1.23), 3.4(1.22), 16.3(1.19), 16.1(1.14), 10.1(1.06), 20.1(1.05), 8.3(1.03), 20.3(1.03), 11.2(1.01), 15.2(1.00), 15.1(0.99), 19.1(0.98), 9.4(0.96), 9.2(0.95), 15.3(0.91), 17.1(0.86), 20.2(0.82), 18.2(0.78), 3.3(0.76), 1.1(0.70), 18.3(0.69)
Biologi	A	Køn	54	3	18.2(0.88), 6.2(0.72), 17.2(0.59)
Biologi	A	Oprindelse	54	7	6.1(1.90), 19.2(1.43), 5.2(1.05), 17.3(0.96), 10.1(0.95), 16.2(0.83), 19.3(0.82)
Biologi	A	Uddannelse	54	4	6.1(1.73), 19.2(1.34), 16.1(0.85), 5.2(0.79)
Biologi	B	Dansk	37	25	1.21(5.15), 1.22(4.97), 2.62(1.46), 2.42(1.45), 1.71(1.35), 2.31(1.34), 3.51(1.13), 3.41(1.13), 3.21(1.12), 2.51(1.12), 3.31(1.11), 1.72(1.05), 3.33(1.03), 2.21(1.02), 1.52(0.98), 3.42(0.97), 2.61(0.94), 1.62(0.94), 3.11(0.91), 2.12(0.90), 3.32(0.85), 2.32(0.84), 2.43(0.83), 2.22(0.76), 3.22(0.65)
Biologi	B	Køn	37	1	1.22(2.44)
Biologi	B	Oprindelse	37	7	1.23(1.69), 2.41(1.37), 3.52(1.28), 3.51(1.12), 3.41(1.02), 2.31(0.90), 2.61(0.88)
Biologi	B	Uddannelse	37	3	1.22(1.97), 2.21(1.19), 3.41(1.09)
Fysik/kemi	A	Dansk	56	32	19.1(1.85), 10.1(1.83), 21.3(1.81), 15.2(1.80), 21.4(1.79), 3.1(1.77), 8.1(1.75), 21.2(1.66), 16.2(1.59), 9.1(1.59), 8.2(1.58), 17.1(1.56), 8.3(1.55), 6.4(1.51), 7.4(1.48), 10.2(1.36), 15.4(1.27), 19.3(1.21), 22.1(1.16), 20.2(1.11), 7.1(1.09), 4.1(1.04), 21.1(1.00), 7.2(0.96), 13.2(0.95), 13.3(0.93), 2.2(0.93), 12.1(0.88), 15.1(0.88), 20.3(0.87), 1.3(0.84), 16.1(0.83)
Fysik/kemi	A	Køn	56	6	5.2(1.02), 18.1(0.99), 15.1(0.96), 17.3(0.90), 14.1(0.90), 12.2(0.70)
Fysik/kemi	A	Oprindelse	56	3	22.2(1.33), 10.2(1.26), 19.3(1.13)
Fysik/kemi	A	Uddannelse	56	7	3.1(1.21), 21.2(1.10), 10.1(0.92), 10.2(0.91), 19.3(0.90), 19.1(0.81), 8.1(0.76)
Fysik/kemi	B	Dansk	42	29	2.13(2.36), 3.42(1.79), 3.21(1.73), 3.51(1.71), 3.41(1.63), 1.42(1.58), 1.62(1.57), 3.31(1.53), 2.41(1.47), 3.62(1.36), 3.11(1.36), 2.26(1.31), 1.72(1.29), 2.31(1.28), 2.24(1.25), 1.22(1.24), 1.31(1.22), 1.51(1.19), 2.32(1.18), 2.52(1.12), 3.61(1.11), 1.52(1.05), 1.43(1.03), 1.71(1.00), 3.32(0.98), 2.23(0.90), 3.12(0.90), 2.54(0.84), 2.42(0.73)
Fysik/kemi	B	Køn	42	3	2.13(3.96), 2.14(3.55), 2.25(0.92)
Fysik/kemi	B	Oprindelse	42	5	2.13(3.15), 2.14(2.64), 2.41(1.35), 3.21(1.29), 3.12(1.23)

Fag	Del	Dimension	Antal items	Antal items med DIF > 0,5	Items med DIF > 0,5
Fysik/kemi	B	Uddannelse	42	14	2.13(3.42), 2.14(2.75), 3.21(1.43), 1.43(1.22), 2.52(1.17), 1.21(1.04), 3.51(0.99), 2.41(0.93), 3.41(0.92), 3.61(0.91), 3.11(0.90), 1.51(0.87), 3.31(0.82), 1.31(0.75)
Geografi	A	Dansk	58	34	5.1(2.67), 20.3(2.03), 18.4(1.99), 14.1(1.75), 2.2(1.71), 11.1(1.68), 1.3(1.57), 13.4(1.57), 21.3(1.53), 15.3(1.50), 19.1(1.47), 18.3(1.43), 18.2(1.26), 14.4(1.24), 7.1(1.20), 2.4(1.14), 21.2(1.13), 3.1(1.08), 13.1(1.06), 1.1(1.03), 7.2(1.03), 6.1(1.03), 20.1(1.02), 20.4(1.02), 3.2(0.98), 13.2(0.97), 20.2(0.90), 19.2(0.87), 15.2(0.82), 8.1(0.73), 14.2(0.73), 4.1(0.65), 10.2(0.62), 14.3(0.59)
Geografi	A	Køn	58	4	14.1(0.94), 4.4(0.85), 14.2(0.67), 10.1(0.65)
Geografi	A	Oprindelse	58	11	20.3(1.38), 2.2(1.22), 1.1(1.17), 2.4(1.00), 20.4(0.94), 18.4(0.92), 15.1(0.90), 20.1(0.89), 13.2(0.82), 3.1(0.77), 4.3(0.74)
Geografi	A	Uddannelse	58	23	11.1(1.66), 20.3(1.33), 18.4(1.17), 21.3(1.15), 2.2(1.12), 14.4(1.08), 2.4(1.01), 15.1(0.97), 19.1(0.96), 15.2(0.95), 3.1(0.93), 13.4(0.93), 14.1(0.84), 14.2(0.80), 1.1(0.77), 18.2(0.76), 20.2(0.74), 18.3(0.72), 1.3(0.71), 6.1(0.69), 2.3(0.68), 13.1(0.64), 20.4(0.64)
Geografi	B	Dansk	37	16	1.41(1.84), 2.11(1.67), 3.22(1.48), 2.23(1.30), 3.42(1.12), 1.31(1.02), 3.41(0.90), 3.62(0.84), 1.11(0.83), 1.52(0.80), 3.72(0.79), 2.41(0.73), 2.52(0.71), 1.53(0.62), 2.31(0.57), 3.61(0.55)
Geografi	B	Køn	37	0	
Geografi	B	Oprindelse	37	9	1.41(1.31), 2.42(1.27), 1.52(1.16), 2.11(1.14), 1.11(1.11), 3.42(1.04), 2.52(0.95), 1.31(0.84), 3.61(0.81)
Geografi	B	Uddannelse	37	10	2.11(1.79), 3.31(1.17), 1.52(1.14), 1.41(1.03), 3.41(1.02), 1.31(1.01), 2.51(0.92), 3.61(0.90), 2.52(0.75), 3.42(0.73)

Anm.: Grupperingen 'Dansk' opdeler børn i to grupper, afhængigt af om deres karaktergennemsnit i dansk læsning og dansk retskrivning er over 7. Køn opdeler eleverne i grupper efter deres biologiske køn (dreng og pige). Oprindelse opdeler eleverne i to grupper efter, om begge deres forældre er født i Danmark. Uddannelse opdeler eleverne i to grupper, afhængigt af om mindst en forælder højst har gennemført grundskolen. Tallene i parentes angiver forskellen i logit på tværs af grupperne.

Kilde: VIVEs analyser på baggrund af data fra STIL og Danmarks Statistik.

Litteratur

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly*, *71*(2), 287–311.

Beckler, D. T., Thumser, Z. C., Schofield, J. S., & Marasco, P. D. (2018). Reliability in evaluator-based tests: Using simulation-constructed models to determine contextually relevant agreement thresholds. *BMC Medical Research Methodology, 18,* 1-12.

Børne- og Undervisningsministeriet (2023). *Generel information om forsøget*. Tilgået d. 25. 1. 2024: https://www.uvm.dk/folkeskolen/folkeskolens-proever/proevetilrettelaeggelse/adgang-tilmelding-og-booking/proeveformer-og-forsoeg/forsoeg-med-de-skriftlige-proever-i-naturfag/generel-information-om-forsoeget.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194.

Dolin, J. (2014). Naturfaglige kompetencer – om kompetencetænkningen i nye Forenklede Fælles Mål. I: S. Tougaard & L. H. Kofod (red.), *Metoder i naturfag – en antologi* (49-66). København: Experimentarium.

Emu (2023). *Skrivning i matematik*. Tilgået d. 25. 1. 2024]: https://emu.dk/grundskole/matematik/sproglig-udvikling/skrivning-i-matematik.

PhET (2023). *Interactive Simulations for Science and Math*. Tilgået d. 25. 1. 2024]: https://phet.colorado.edu/.

Rambøll (2018). *Statusnotat. Evaluering og følgeforskning. Indførelse af den ny fælles prøve i fysik/kemi, biologi og geografi – prøvens betydning for undervisningens form og indhold*. København: Rambøll.

Regeringen (2018). *National naturvidenskabsstrategi.* København: Undervisningsministeriet.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273-1296.

Wright, B. D., & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. *Rasch Measurement Transactions*, *8*, 370–371.