

Evaluering af forsøgsprøver – biologi, fysik/kemi og geografi

Hovedresultater



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 udtræksprøverne 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 udtræksprøver, 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 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 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/proevetilrettelaeg-gelse/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 andel 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 tekstsvare 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.

Hovedresultater: Evaluering af forsøgsprøver – biologi, fysik/kemi og geografi

VIVE/2024

HR_302410