BØRNE- OG
UNDERVISNINGS-
MINISTERIET
STYRELSEN
FOR IT OG LÆRING

Genberegning af opgavernes sværhedsgrad, elevernes dygtighed og testenes statistiske usikkerhed

Genberegning af opgavernes
sværhedsgrad, elevernes dygtighed og testenes statistiske usikkerhed

Indhold

Sammenfatning

Politisk aftale om justering af de nationale test:

Der blev i februar 2020 indgået aftale i folkeskoleforligskredsen om nationale test. Det følger af aftalen, at der skal gennemføres mindre forbedringer af det eksisterende testsystem i løbet af 2020, så de obligatoriske test kan gennemføres fra skoleåret 2020/2021, og frem til et nyt evaluerings- og bedømmelsessystem træder i kraft. Forbedringerne skal gøre testene mere retvisende på elevniveau.

I juni 2020 besluttede folkeskoleforligskredsen, at forbedringerne skal ske ved, at tilbagemeldingen på profilområder slås sammen, så der fremover gives én samlet tilbagemelding.

Formålet er at reducere den statistiske usikkerhed på elevniveau.

- Opgaverne i hver af en tests nuværende tre adskilte profilområder er samlet i én opgavebank med tilhørende sværhedsgrader, og opgavernes sværhedsgrader er genberegnet.
- Elevernes dygtighed og den tilhørende statistiske usikkerhed er genberegnet ved anvendelse af opgavernes genberegnede sværhedsgrader.

- Genberegningen af opgavernes sværhedsgrad har betydet, at 1,2 procent af opgaverne ikke kan indplaceres i den nye opgavebank, hvorfor disse slettes og ikke vil blive anvendt fremover.
- Der er en stor sammenhæng mellem opgavernes tidligere og nye sværhedsgrader.

- Den statistiske usikkerhed på den samlede elevdygtighed i hver test er i gennemsnit faldet til 0,27. Usikkerheden på de oprindelige elevdygtigheder i hvert profilområde lå i gennemsnit på 0,46.
- Den statistiske usikkerhed på den samlede elevdygtighed er under 0,30 for 72 procent af testresultaterne og under 0,35 for 95 procent af testresultaterne.
- Den statistiske sikkerhed kan yderligere forbedres ved at forlænge testtiden samt ved at fokusere på bestemte opgavetyper i forbindelse med opgaveproduktionen.

- Der er en stor sammenhæng mellem elevernes oprindelige og de genberegnede dygtigheder. Korrelationerne ligger mellem 0,74 og 0,95.

- På landsplan er udviklingen i resultaterne i de nationale test i dansk, læsning, i matematik og i fysik/kemi næsten den samme på den normbaserede percentilskala (1-100) som på den nye pointskala (0-50). Den nuværende percentilskala til formid-

ling af resultaterne har flere uheldige statistiske egenskaber, hvorfor denne er udskiftet med pointskalaen. I engelsk 7. klasse viser de genberegnede elevdygtigheder opgjort på pointskalaen en lille stigning over årene. De oprindelige resultater på percentilskalaen viste et lille fald i 2015/16, hvorefter landsresultatet næsten har været konstant.

- Den kriteriebaserede skala er justeret som følge af ændringen i beregningen af elevdygtigheden.
- På landsplan er der små og få ændringer i opgørelserne af de nationale måltal, der er baseret på den kriteriebaserede tilbagemelding. Udviklingen i de nationale måltal er stort set uændret efter genberegningerne.

- Reliabiliteten, dvs. testens evne til at nå det samme resultat ved gentagne målinger, ligger i intervallet 0,85-0,95. Reliabiliteten er forbedret i forhold til tidligere.
- Der er en tydelig sammenhæng mellem elevernes resultater fra de nationale test i dansk, læsning og matematik i 8. klasse og elevernes karakterer i folkeskolens prøver i 9. klasse, hvilket er en indikation af, at de nationale test måler det samme som folkeskolens prøver i sammenlignelige fag. Validiteten af testene er på niveau med tidligere.

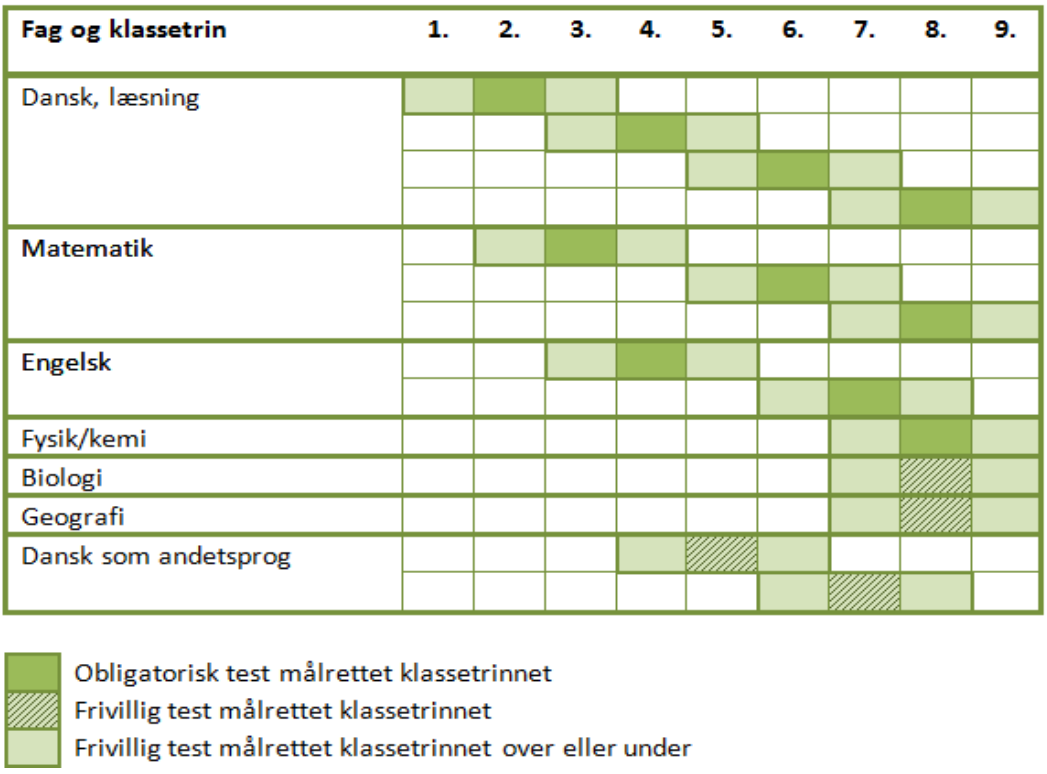
1 Indledning

1.1 De nationale test

De nationale test er it-baserede, selvscorende og adaptive. At testene er adaptive betyder, at opgaverne i et testforløb udvælges, så de bedst muligt passer til elevens dygtighedsniveau undervejs i forløbet. Dygtige elever får de sværeste opgaver, mens elever med større faglige udfordringer får de lettere opgaver.

Der er ti obligatoriske nationale test i folkeskolen og yderligere fire frivillige nationale test (figur 1), hvor hver test består af tre faglige profilområder[1]. En test kan gennemføres på 45 minutter.

Figur 1 Frivillige og obligatoriske nationale test



Kilde: www.uvm.dk/folkeskolen/elevplaner-nationale-test--trivselsmaaling-og-sprogproever/nationale-test

De første obligatoriske nationale test blev gennemført i folkeskolen i skoleåret 2009/2010.

[1] https://www.uvm.dk/folkeskolen/elevplaner-nationale-test--trivselsmaaling-og-sprogproever/nationale-test/klassetrin-fag-og-profilomraader

I hvert fag testes eleverne i tre faglige hovedområder, der kaldes profilområder.

Profilområder:
- **Dansk/læsning**: (1) Sprogforståelse, (2) afkodning og (3) tekstforståelse
- **Matematik**: (1) Tal og algebra, (2) geometri og (4) statistik og sandsynlighed
- **Engelsk**: (1) Læsning, (2) ordforråd og (4) lytning (4. kl.) / (3) sprog og sprogbrug (7. kl.)
- **Fysik/kemi**: (1) Energi og energiomsætning, (2) fænomener, stoffer og materialer og (3) anvendelser og perspektiver
- **Biologi**: (1) Den levende organisme, (2) Levende organismers samspil med hinanden og deres omgivelser og (3) biologiens anvendelse, tankegange og arbejdsmetoder
- **Geografik**: (1) Naturgrundlaget, (2) kulturgeografi og (3) at bruge geografien
- **Dansk som andetsprog**: (1) Ordforråd, (2) sprog og sprogbrug og (3) læseforståelse

Der blev i februar 2020 indgået aftale i folkeskoleforligskredsen om nationale test, hvoraf det fremgår, at der skal gennemføres mindre forbedringer af det eksisterende testsystem[2]. Forbedringerne skal gøre testene mere retvisende på elevniveau.

I juni 2020 besluttede folkeskoleforligskredsen[3], at forbedringerne skal ske ved, at tilbagemeldingen på de tre profilområder i hver test slås sammen, så der fremover gives én samlet tilbagemelding. Formålet er at reducere den statistiske usikkerhed på elevniveau.

Beregning af én samlet elevdygtighed med tilhørende statistisk usikkerhed i hver test kræver, at opgaverne i hver af en tests nuværende tre adskilte profilområder samles i én opgavebank med tilhørende sværhedsgrader.

Der er derfor brug for at genberegne opgavernes sværhedsgrader. Genberegningen af opgavernes sværhedsgrad sikrer, at opgaverne i opgavebanken knyttet til hver test er korrekt indplaceret i forhold til deres sværhedsgrad.

Opgavernes genberegnede sværhedsgrader anvendes fra og med foråret 2021 i de nationale test.

For endvidere at kunne foretage sammenligninger med tidligere resultater på elev, skole, kommune og nationalt niveau er elevernes dygtigheder genberegnet tilbage i tid med anvendelse af opgavernes genberegnede sværhedsgrader.

Dette notat viser resultaterne af genberegningerne af opgavernes sværhedsgrader samt af elevdygtighederne og den tilhørende statistiske usikkerhed. Desuden præsenteres sammenligninger af de oprindeligt beregnede elevdygtigheder og de genberegnede. Testenes reliabilitet og validitet er ligeledes beregnet.

[2] https://www.uvm.dk/aktuelt/nyheder/uvm/2020/feb/200221-ny-aftale-skaber-tryghed-om-kvaliteten-af-elevernes-bedoemmelsesresultater

[3] Forslag til lov om ændring af lov om folkeskolen (Gennemførelse af nationale test i skoleåret 2019/2020) (Lovforslag nr. L 126)

Kort om Rasch-modellen til beregning af elevdygtighed i nationale test

De nationale test er baseret på Rasch-modellen. Rasch-modellen er en sandsynlighedsmodel, der i den simpleste udgave, kaldet det dikotome tilfælde, giver sandsynligheden for, at elev nummer *n* med dygtighedsparameteren θ_n svarer rigtigt (svarende til scoringen $X_{ni}=1$) på opgave nummer *i* med sværhedsparameteren β_i:

$$P\{X_{ni} = x\} = \frac{e^{x(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}, \qquad x \in [0,1]$$

Sandsynligheden for, at en elev svarer rigtigt på en opgave, afhænger således kun af elevens dygtighed θ (theta) og opgavens sværhed β (beta).

I Rasch-modellen optræder opgavesværheder og elevdygtigheder på samme skala.

Rasch-modellen kan udvides til at inkludere polytome opgaver, dvs. opgaver med flere delopgaver, således at scoringen x kan antage højere heltalsværdier end 1 svarende til, at flere delspørgsmål er besvaret korrekt.

I de nationale test anvendes både dikotome og polytome opgaver.

Opgavernes sværhedsgrad er indtil foråret 2021 beregnet på baggrund af elevbesvarelser i deciderede opgaveafprøvninger.

Ved anvendelse af opgavernes sværhedsgrader kan elevernes dygtighed beregnes.

Desuden beregnes den statistiske usikkerhed, ***S**tandard **E**rror of the **M**easurement* (SEM), på elevdygtigheden.

1.2 Resultatskalaer

I de nationale test anvendes forskellige skalaer til såvel beregning som til formidling af testresultaterne.

1.2.1 Logit-skala

Elevernes dygtighed beregnes på logit-skalaen[4]. Logit-skalaen er en intervalskala fra minus til plus uendelig, hvor værdierne i praksis ligger mellem -7 og +7. Formidling af resultater på denne skala kan være forbundet med vanskeligheder, hvorfor formidlingen af testresultaterne til elev, lærer, forældre, skoleleder, den kommunale forvaltning samt på nationalt niveau fra opstarten af de nationale test i 2010 til og med 2020 er foregået på den normbaserede percentilskala, som går fra 1 til 100.

1.2.2 Percentilskala (1-100)

Normen i den normbaserede percentilskala henviser til, at skalaen tager udgangspunkt i første gang, der blev gennemført nationale test.

[4] Logit er en transformation med den naturlige logaritme af odds, p/(1-p), hvor *p* er sandsynligheden for at svare rigtigt på et item.

Princippet bag percentilskalaen er, at de 1 procent af eleverne med de dårligste resultater får en score på 1, de næste 1 procent af eleverne med de næstdårligste resultater får scoren 2 osv. op til, at de 1 procent af eleverne med de bedste resultater får scoren 100.

Percentilskalaen har flere uheldige statistiske egenskaber. Blandt andet fremstår forskelle mellem elevdygtigheder på logit-skalaen forskelligt, når disse omregnes til percentilskalaen. Forskellen afhænger af elevernes dygtighedsniveau på logit-skalaen.

1.2.3 Pointskala (0-50)

I forbindelse med justeringen af de nationale test i skoleåret 2020/21 udskiftes den normbaserede percentilskalaen med en normbaseret pointskala (0-50) for at sikre en mere korrekt fortolkning og formidling af resultaterne.

Omregningen af elevdygtighederne fra logit-skalaen til pointskalaen[5] foregår ved anvendelse af en lineær funktion, hvor der er valgt en skalering, så skalaen i skoleåret 2018/19 har et gennemsnit lig 25 og en spredning mellem elevdygtighederne på 5[6], dvs.

$$P = 5 \cdot \tilde{\theta} + 25,$$

hvor $\tilde{\theta}$ er den standardiserede elevdygtighed på logit-skalaen.

Testresultaterne på pointskalaen vil yderligere blive formidlet på en femtrinsskala inddelt efter elevernes dygtigheder[7]:

1. Klart under middel
2. Under middel
3. Middel
4. Over middel
5. Klart over middel

En tilsvarende inddeling i fem trin har også fundet sted i forbindelse med formidling af resultaterne på percentilskalaen fra 2010 til 2020.

1.2.4 Den kriteriebaserede skala

I dansk, læsning og matematik formidles testresultaterne endvidere på den kriteriebaserede skala på seks niveauer. Her vises resultatet som et udtryk for elevernes faglige niveau i de dele af fagene, som testes. De seks niveauer på den kriteriebaserede skala er:

- Fremragende præstation
- Rigtig god præstation
- God præstation

[5] Tan, X. & Michel, R.: Why do Standardized Testing programs Report Scaled Scores? Why not just report the raw or percent-correct scores? Educational Testing Services: R&D Connections no 16., USA, 2011.

[6] I de norske *nasjonale prøver* er valgt et gennemsnit på 50 og en spredning på 10. Utdanningsdirektoratet 2018. Metodegrunnlag for nasjonale prøver

[7] Skæringspunkterne er baseret på fordelingen af elevernes dygtigheder i 2018/19 således, at 10 procent af eleverne med de dårligste testresultater i 2018/19 ligger på trin 1, de 25 procent af eleverne med de næstdårligste testresultater ligger på trin 2, 30 procent af eleverne ligger på trin 3, 25 procent af eleverne ligger på trin 4, og de 10 procent af eleverne med de bedste testresultater ligger på trin 5.

- Jævn præstation
- Mangelfuld præstation
- Ikke tilstrækkelig præstation

De kriteriebaserede tilbagemeldinger indgår i de nationale resultatmål til opfølgning på folkeskolereformen fra 2014:

- Mindst 80 pct. af eleverne skal være gode til at læse og regne i de nationale test.
- Andelen af de allerdygtigste elever i dansk og matematik skal stige år for år.
- Andelen af elever med dårlige resultater i dansk og matematik skal reduceres år for år.

Tabel 1. De seks faglige niveauer og deres sammenhæng til de nationale resultatmål

<table>
<tr><th>Niveau på skalaen</th><th colspan="2">Nationale resultatmål</th></tr>
<tr><td>Fremragende præstation</td><td>Andelen af de allerdygtigste elever skal øges år for år.</td><td rowspan="3">Mindst 80 pct. af eleverne skal være gode til at læse og regne.</td></tr>
<tr><td>Rigtig god præstation</td><td rowspan="2"></td></tr>
<tr><td>God præstation</td></tr>
<tr><td>Jævn præstation</td><td colspan="2"></td></tr>
<tr><td>Mangelfuld præstation</td><td colspan="2" rowspan="2">Andelen af elever med dårlige resultater skal reduceres år for år.</td></tr>
<tr><td>Ikke tilstrækkelig præstation</td></tr>
</table>

Kilde: Styrelsen for it og læring

I forbindelse med justeringen af de nationale test, hvor de nuværende beregninger af elevdygtigheden i hver af de tre profilområder i hver test bliver erstattet med én samlet beregning af elevdygtigheden, er det også nødvendigt at justere den kriteriebaserede skala.

Grænseværdierne, kaldet *cut-scores*, på den kriteriebaserede skala er fra og med skoleåret 2020/21 fastsat således, at elevfordelingen på de enkelte kriterier i størst muligt omfang er bevaret efter samlingen af elevbesvarelserne i de tre profilområder. Fastsættelsen af de nye grænseværdier tager udgangspunkt i fordelingen af elevdygtighederne på den kriteriebaserede skala i 2018/19.

1.3 Opgaveafprøvning og fastsættelse af opgavernes sværhedsgrad

Opgaverne, der anvendes i de nationale test, er udarbejdet af faglige opgavekommissioner. Alle opgaver i opgavebanken er afprøvet af elever på det klassetrin, testen er målrettet til. Opgaverne er afprøvet af ca. 700 elever.

På baggrund af elevernes besvarelser fra opgaveafprøvningerne er der foretaget statistisk analyse, hvor det er undersøgt, om opgaverne passer til Rasch-modellen[8,9]. Opgaver, der

[8] Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests. Danish National Institute for Educational Research, Mesa Press. Copenhagen 1960.

[9] Christensen K.B., Kreiner S., Mesbah M. (editors): Rasch Models in Health. John Wiley & Sons, Inc., USA. 2013.

ikke passede til Rasch-modellen blev fjernet (misfit)[10]. Opgaverne er også undersøgt for bias med hensyn til køn, skolestørrelse samt geografi i en ***D**ifferential **I**tem **F**unctioning* (DIF) analyse[11,12,13]. DIF-testen har til formål at sikre, at opgaverne tester eleverne ens ('retfærdigt') uanset køn, skolestørrelse eller skolens geografiske placering. Det vil sige, at testene ikke favoriserer således, at opgaver kræver viden, der er knyttet til en bestemt gruppe elever. Opgaver, der udviste statistisk signifikant DIF, er fjernet. Til de resterende opgaver beregnes opgavernes sværhedsgrad baseret på elevernes besvarelser i opgaveafprøvningen, og opgaverne tilføjes opgavebanken.

Analyserne er foretaget særskilt for hvert profilområde. Opgavernes oprindelige sværhedsgrad er således knyttet til skalaen i det enkelte profilområde.

Alle analyser af besvarelser fra opgaveafprøvningerne er foretaget i analyseprogrammet RUMM[14].

1.4 Datagrundlaget for genberegningen af opgavernes sværhedsgrad

Beregning af én samlet elevdygtighed på tværs af de oprindelige tre profilområder med tilhørende statistisk usikkerhed i hver test kræver, at opgaverne i hver af en tests nuværende tre adskilte profilområder samles i én opgavebank med tilhørende sværhedsgrader på én fælles skala.

Der er derfor brug for at genberegne opgavernes sværhedsgrader. Genberegningen af opgavernes sværhedsgrad sikrer, at opgaverne i opgavebanken knyttet til hver test er korrekt indplaceret i forhold til deres sværhedsgrad.

Til genberegning af opgavernes sværhedsgrad anvendes data fra elevernes besvarelser fra de obligatoriske[15] nationale test i perioden 2015-2019[16]. Disse data består af testforløb, hvor eleverne har besvaret opgaver fra alle tre profilområder i den enkelte test. Det er derfor muligt, at beregne opgavernes sværhedsgrader på tværs af profilområderne.

---

[10] https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf19/mar/190315-opgaveafprovning-og-beregning-af-opgavernes-svarhedsgrad-i-de-nationale-test.pdf

[11] Martinková, P., Drabinová, A ., Liaw, Y.-L., Sanders, E.A., McFarland, J.L., Price, R.M., Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments, CBE Life Sciences Education, 2017, 16(2). www.ncbi.nlm.nih.gov/pmc/articles/PMC5459266/

[12] Hagquist, C., Explaining differential item functioning focusing on the crucial role of external information- an example from the measurement of adolescent mental health, BMC Medical Research Methodology, 2019, 19:185. Doi.org/10.1186/s12874-019-0828-3

[13] Andrich, D., Hagquist, C., Real and Artificial Differential Item Functioning in Polytomous Items, Educational and Psychological Measurement, 2015, 75(2), p. 185–207.

[14] www.rummlab.com.au

[15] For opgaverne i dansk som andetsprog er der anvendt elevbesvarelser fra de frivillige test, da disse test ikke er obligatoriske.

[16] For opgaverne i matematik 6. klasse er der kun anvendt data fra elevbesvarelserne i perioden 2018-2019, da der har været udskiftning af ét af profilområderne. For opgaverne i matematik 8. klasse og engelsk 4. klasse er der kun anvendt data fra elevbesvarelserne i perioden 2018-2019, da disse test først er indført i 2018.

Datagrundlaget til genberegning af opgavernes sværhedsgrader adskiller sig fra datagrundlaget fra tidligere opgaveafprøvninger på to områder:

1. Data stammer fra elevbesvarelser i et adaptivt testforløb. I tidligere opgaveafprøvninger fik alle elever, der deltog, de samme opgaver uanset elevens dygtighed og opgavens sværhedsgrad. I et adaptivt forløb er de letteste opgaver primært besvaret af de mindre dygtige elever, og de sværeste opgaver er primært besvaret af de dygtigste elever.
2. Data stammer fra elevers besvarelser i et egentligt testforløb, hvor eleven er informeret om, at testen afvikles med henblik på en vurdering af elevens dygtighed. I tidligere opgaveafprøvninger blev eleverne informeret om, at testen blev afviklet med henblik på at vurdere opgavernes sværhedsgrad. Eleverne fik ingen tilbagemelding på en opgaveafprøvning.

2 Opgavernes nye sværhedsgrad og undersøgelse af fælles skala

Ved anvendelse af data fra elevernes besvarelser fra de obligatoriske nationale test er samtlige opgaver i opgavebanken analyseret på ny med henblik på at undersøge, hvorvidt opgaverne fra de oprindelige adskilte profilområder passer til én samlet opgavebank i hver test.

Da opgaverne tidligere er undersøgt for bias (DIF) i forhold til køn, skolestørrelse og geografi bliver denne analyse ikke gentaget.

2.1 Opgaver, der ikke passer til opgavebanken

Opgaver, der ikke passede til den fælles Rasch-model i hver test, blev fjernet fra opgavebanken. I alt blev 141 opgaver (1,2 procent af den samlede opgavebank), som ikke længere passer til Rasch-modellen, kasseret (tabel 2).

For hovedparten af testene er det under 1 procent af det samlede antal opgaver, som ikke passer ind i opgavebanken efter genberegningen. I opgavebanken til dansk, læsning i 2. klasse og til engelsk i 4. klasse er der en større andel opgaver (knap 4 procent), som ikke passer til Rasch-modellen på den fælles skala.

I bilag 1 i bilagsnotatet findes en uddybende tabel på tværs af de oprindelige profilområder.

Tabel 2. Antal opgaver i opgavebanken før og efter samling af profilområder og genberegning af sværhedsgrader

Fag	Klassetrin	Antal opgaver i opgavebanken 2015-20	Antal opgaver i opgavebanken 2021-	Antal opgaver, som udgår
Dansk, læsning	2. klasse	763	733	30 (3,9 pct.)
	4. klasse	818	815	3 (0,4 pct.)
	6. klasse	777	773	4 (0,5 pct.)
	8. klasse	826	824	2 (0,2 pct.)
Matematik	3. klasse	755	754	1 (0,1 pct.)
	6. klasse	1.100	1.096	4 (0,4 pct.)
	8. klasse	707	705	2 (0,3 pct.)
Engelsk	4. klasse	788	760	28 (3,6 pct.)
	7. klasse	952	934	18 (1,9 pct.)
Fysik/kemi	8. klasse	970	968	2 (0,2 pct.)
Biologi	8. klasse	784	772	12 (1,5 pct.)
Geografi	8. klasse	982	977	5 (0,5 pct.)
Dansk som andetsprog	5. klasse	812	799	13 (1,6 pct.)
	7. klasse	982	965	17 (1,7 pct.)
Samlet		**12.016**	**11.875**	**141 (1,2 pct.)**

Kilde: Styrelsen for It og læring

Opgaver, der ikke passer til Rasch-modellen efter genberegningen, fjernes fra opgavebanken og indgår ikke i den efterfølgende genberegning af elevdygtighederne.

Analyserne af opgaverne har således vist, at det er muligt at samle langt hoveparten (98,8 procent) af opgaverne fra de oprindelige tre profilområder til én samlet opgavebank i hver test.

2.2 Analyse af én-dimensionalitet

For at undersøge om den samlede elevdygtighed på tværs af de oprindelige tre profilområder er målt på én og samme skala, er korrelationerne mellem den samlede elevdygtighed i hver test og elevdygtighederne i de tre adskilte profilområder beregnet. Genberegningen af elevdygtighederne i hver af de enkelte profilområder er foretaget ved anvendelse af opgavernes genberegnede sværhedsgrader, men hvor der kun medtages elevbesvarelser af opgaver fra det enkelte profilområde.

Korrelationerne ligger i intervallet 0,74-0,92 (tabel 3).

Desuden blev egenværdierne for korrelationsmatricen beregnet med henblik på at finde andel varians for den største egenværdi (*Den Principale Komponent*) ud af den samlede varians for alle tre egenværdier. Denne andel kan betragtes som et mål for den 1-dimensionalitet, der kan findes blandt testens oprindelige tre profilområder. For alle test viste

beregningerne en bevarelse af den samlede varians ved sammenlægning af profilområderne på mindst 68 procent, således at sammenlægningen bevarer informationen fra de tre oprindelige profilområder.

Tabel 3. Korrelation mellem genberegnet samlet elevdygtighed og elevdygtighed i hvert af de oprindelige tre profilområder

Fag	Klassetrin	Korrelation mellem genberegnet elevdygtigheder på fælles skala og i hvert profilområde		
		Profilområde 1	Profilområde 2	Profilområde 3
Dansk, læsning	2. klasse	0,76	0,89	0,92
	4. klasse	0,84	0,86	0,89
	6. klasse	0,81	0,84	0,86
	8. klasse	0,74	0,85	0,82
Matematik	3. klasse	0,88	0,85	0,89
	6. klasse	0,88	0,87	0,88
	8. klasse	0,92	0,92	0,91
Engelsk	4. klasse	0,95	0,93	0,88
	7. klasse	0,92	0,92	0,93
Fysik/kemi	8. klasse	0,81	0,80	0,81
Biologi	8. klasse	0,81	0,83	0,79
Geografi	8. klasse	0,82	0,83	0,84
Dansk som andetsprog	5. klasse	0,88	0,89	0,90
	7. klasse	0,88	0,88	0,88

Note: Pearsons korrelationskoefficient.
Kilde: Styrelsen for It og læring

For endvidere at sikre, at principal komponent-analysen kan bruges i forbindelse med sammenlægningen af de tre profilområder, blev korrelation mellem elevdygtighederne på Rasch-skalaen og principal komponent-skalaen beregnet. Denne analyse viser en høj grad af sammenhæng mellem Rasch- og principal komponent-skalaen med korrelationer over 0,98 for samtlige test.

3 Sammenligning mellem opgavernes oprindelige og de genberegnede sværhedsgrader

Opgavernes genberegnede sværhedsgrader kan ikke direkte sammenlignes med opgavernes tidligere sværhedsgrader. Dette skyldes, at parameteriseringen af opgaverne afhænger af de øvrige opgavers sværhedsgrad i samme opgavebank. Opgavernes tidligere sværhedsgrader var beregnet i forhold til opgaverne i hvert af de tre selvstændige profilområder.

Derimod er det muligt at sammenligne opgavernes indbyrdes rangorden i hvert af de tre *oprindelige* profilområder med opgavernes rangorden i forhold til deres *genberegnede* sværhedsgrader i den samlede opgavebank for hver test.

3.1 Korrelationer mellem opgavernes sværhedsgrader

For at vurdere, hvorvidt rangordenen er intakt blandt opgaverne, når profilområderne er slået sammen, og opgavernes sværhedsgrader er genberegnet, undersøges korrelationen mellem opgavernes oprindelige sværhedsgrader i de adskilte profilområder og opgavernes genberegnede sværhedsgrader på den nye fælles skala.

Korrelationskoefficienterne i tabel 4 vidner om en høj korrelation på tværs af alle fag og klassetrin. Det samme gælder for korrelationskoefficienterne i de enkelte profilområder (bilag 2 i bilagsnotatet). Sammenhængen mellem tidligere og genberegnede sværhedsgrader er også illustreret ved *scatterplots* i bilag 3 i bilagsnotatet.

Tabel 4. Korrelationskoefficienter mellem tidligere og de genberegnede opgavesværhedsgrader på tværs af fag og klassetrin

Fag	Klassetrin	Korrelation
Dansk, læsning	2. klasse	0,86
	4. klasse	0,82
	6. klasse	0,75
	8. klasse	0,78
Matematik	3. klasse	0,88
	6. klasse	0,92
	8. klasse	0,96
Engelsk	4. klasse	0,84
	7. klasse	0,88
Fysik/kemi	8. klasse	0,91
Biologi	8. klasse	0,91
Geografi	8. klasse	0,89
Dansk som andetsprog	5. klasse	0,89
	7. klasse	0,84
Samlet		**0,86**

Note: Pearsons korrelationskoefficient.
Kilde: Styrelsen for It og Læring

Korrelationskoefficienterne er større end 0,80 i 12 ud af de 14 test. I dansk, læsning på 6. og 8. klassetrin ligger korrelationen på henholdsvis 0,75 og 0,78.

De høje korrelationskoefficienter vidner om, at rangordenen mellem opgavernes sværhedsgrader i vid udstrækning er bibeholdt efter, at opgaverne fra de tidligere separate opgavebanker knyttet til hver af en tests tre profilområder er samlet til én fælles opgavebank for hver test, og opgavernes sværhedsgrader herefter er genberegnet.

4 Den statistisk usikkerhed på de genberegnede elevdygtigheder

Elevernes dygtighed og den tilhørende statistiske usikkerhed, SEM, er genberegnet for de obligatoriske nationale test afholdt i perioden 2015-2020. For testen i dansk som andetsprog er der udelukkende tale om frivillige nationale test, hvorfor genberegningerne i denne test vedrører de frivillige nationale test.

Til genberegningerne er anvendt elevernes besvarelser fra de samlede testforløb i de tre oprindelige profilområder sammen med opgavernes genberegnede sværhedsgrader.

Genberegningen af elevdygtighederne anvender kun de opgaver, der passer ind på den nye skala (tabel 2).

4.1 Usikkerheden på elevdygtighederne

I perioden 2015-2020 har den gennemsnitlige statistiske usikkerhed ligget på 0,46 på logit-skalaen. I de genberegnede elevforløb, som nu er baseret på elevernes besvarelser i alle tre profilområder, er *den gennemsnitlige statistiske usikkerhed reduceret til 0,27* (tabel 5). Der er således tale om en reduktion på over 40 procent.

Tabel 5. Den statistiske usikkerhed (SEM) på elevdygtighederne opdelt på fag, klassetrin og profilområde (median) samt antallet af opgaver i hvert elevforløb (gennemsnit)

Fag	Klassetrin	De oprindelige SEM for hvert profilområde			SEM genberegnet	Antal opgaver	
		Profilområde 1	Profilområde 2	Profilområde 3		Oprindelige elevforløb	Genberegnet elevforløb
Dansk, læsning	2. klasse	0,46	0,47	0,42	0,28	72	64
	4. klasse	0,48	0,48	0,45	0,28	65	64
	6. klasse	0,48	0,49	0,46	0,29	61	59
	8. klasse	0,50	0,51	0,49	0,29	57	56
Matematik	3. klasse	0,47	0,50	0,51	0,29	52	52
	6. klasse	0,47	0,49	0,50	0,30	48	48
	8. klasse	0,52	0,52	0,52	0,31	51	50
Engelsk	4. klasse	0,40	0,42	0,43	0,26	68	59
	7. klasse	0,45	0,47	0,45	0,29	55	50
Fysik/kemi	8. klasse	0,35	0,35	0,33	0,20	57	57
Biologi	8. klasse	0,37	0,35	0,41	0,23	57	55
Geografi	8. klasse	0,35	0,36	0,36	0,22	58	58
Dansk som andetsprog	5. klasse	0,46	0,46	0,45	0,28	50	50
	7. klasse	0,49	0,46	0,50	0,30	53	50

Kilde: Styrelsen for It og Læring

Af beregningerne fremgår det (tabel 5), at de genberegnede elevdygtigheder på tværs af fag og klassetrin har en væsentlig mindre statistisk usikkerhed sammenlignet med de oprindelige.

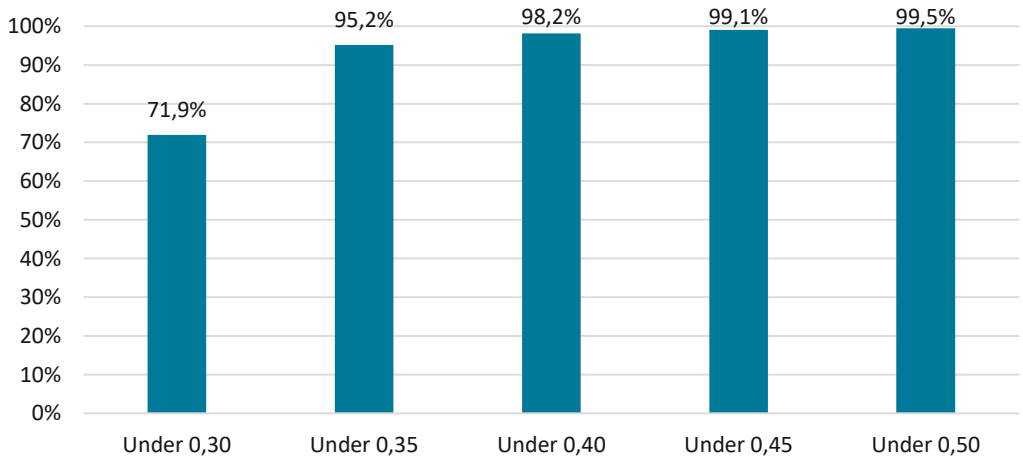
Det samme gælder inden for hvert skoleår, jf. bilag 4 i bilagsnotatet, hvor det fremgår, at den beregnede statistiske usikkerhed på elevdygtighederne stort set er den samme i alle årene.

Rådgivningsgruppen for evaluering af de nationale test anbefalede et niveau for den statistiske usikkerhed på elevdygtigheden på 0,30, hvis testen skulle anvendes i det pædagogiske arbejde. Medianen er under 0,30 i alle test (tabel 5) med undtagelse af i matematik for 8. klassetrin, hvor den er på 0,31.

Tabel 5 viser desuden, hvor mange opgaver beregningerne af den statistiske usikkerhed er baseret på. I snit bygger de nuværende beregninger af elevdygtigheder og tilhørende statistisk usikkerhed på 59 opgaver pr. elev pr. test på tværs af de tre profilområder, mens de genberegnede elevdygtigheder og tilhørende usikkerheder i snit bygger på 57 opgaver pr. elev pr. test. Flere af opgaverne er polytome opgaver, dvs. opgaver der indeholder flere delopgaver. Dette er specielt kendetegnet for testen i fysik/kemi, geografi og biologi, hvor det også fremgår af tabel 5, at den statistiske usikkerhed på elevdygtigheden er mindst.

71,9 procent af eleverne har en beregnet statistisk usikkerhed på elevdygtigheden på under 0,30, jf. figur 2, i det samlede genberegnede elevforløb. Og 95,2 procent af alle elever har efter genberegningen en statistisk usikkerhed på elevdygtigheden på under 0,35 på logit-skalaen.

Figur 2 Fordeling af den statistiske usikkerhed på elevniveau efter genberegning af elevdygtigheder



Note: Diagrammet viser den kumulative frekvens for den statistiske usikkerhed på elevniveau for de genberegnede elevdygtigheder.
Kilde: Styrelsen for It og Læring

For alle fag på alle klassetrin gælder det, at over 90 procent af eleverne har en statistisk usikkerhed på elevdygtigheden på mindre end 0,35 – dog med undtagelse af matematik

på 8. klassetrin, hvor kun 87,8 procent af eleverne har en statistisk usikkerhed under 0,35 (tabel 6).

Tabel 6. Den kumulative frekvens (%) for usikkerheden på elevniveau på tværs af fag og klassetrin

		Den kumulative frekvens, SEM <				
Fag	**Klassetrin**	**< 0,30**	**< 0,35**	**< 0,40**	**< 0,45**	**< 0,50**
Dansk, læsning	2. kl.	68,2	94,0	97,3	98,7	99,3
	4. kl.	75,3	97,3	98,9	99,5	99,7
	6. kl.	67,8	97,3	99,1	99,6	99,8
	8. kl.	68,1	96,2	98,7	99,6	99,8
Matematik	3. kl.	64,8	91,2	95,8	97,8	98,8
	6. kl.	56,6	92,1	96,7	98,3	99,1
	8. kl.	40,7	87,8	94,5	97,4	98,7
Engelsk	4. kl.	72,6	93,3	98,4	99,3	99,6
	7. kl.	62,6	91,9	98,2	99,2	99,6
Fysik/kemi	8. kl.	94,1	98,9	99,5	99,7	99,8
Biologi	8. kl.	89,2	98,4	99,4	99,7	99,8
Geografi	8. kl.	91,4	98,5	99,3	99,6	99,7
Dansk som andetsprog	5. kl.	70,9	94,8	97,4	98,7	99,2
	7. kl.	58,1	92,5	97,3	98,5	99,2
Samlet		**71,9**	**95,2**	**98,2**	**99,1**	**99,5**

Note: Antallet af opgaver svarer til det gennemsnitlige antal opgaver pr. elev pr. test, som beregningen af usikkerhed bygger på. Tidligere opgaver er summen af antal opgaver i de tre profilområder, mens det genberegnede antal opgaver er de opgaver, som genberegningen af elevdygtighed bygger på. Da der er fjernet opgaver, der ikke passede på den nye fælles skala, er antal opgaver i de genberegnede forløb lavere end antal opgaver i de oprindelige forløb.
Kilde: Styrelsen for It og Læring

28,1 procent af eleverne har en beregnet statistisk usikkerhed på elevdygtigheden over 0,30. Dette kan skyldes flere forhold. Dels anvender genberegningen af elevdygtighederne og den tilhørende statistiske usikkerhed ikke de opgaver der, i forbindelse med samlingen af de tre profilområder til én fælles skala, blev fjernet fra opgavebanken, dels er der elever, hvor der kræves besvarelse af flere opgaver, end der er mulighed for på 45 minutter, for at opnå en større sikkerhed.

Andelen af testresultater, hvor SEM er større end 0,30, er størst i matematik i 6. og 8. klasse. Her har henholdsvis 43,4 procent og 59,3 procent af eleverne en beregnet dygtighed, hvor den statistiske usikkerhed er over 0,30. Dette hænger sammen med, at antallet af opgaver, eleverne kan nå at besvare, generelt er lavere i disse test (tabel 5).

Uanset fag og klassetrin gælder, at usikkerheden på testresultaterne er størst for elever med de bedste testresultater. Således er usikkerheden over 0,30 blandt 50 procent af de elever, hvor den beregnede dygtighed ligger klart over middel.

Den statistiske usikkerhed på elevernes beregnede dygtighed kan nedbringes yderligere primært ved, at eleverne besvarer flere opgaver, hvilket kræver længere testtid. Desuden kan usikkerheden nedbringes ved at supplere opgavebanken med flere opgaver således,

at der er tilstrækkeligt med opgaver med sværhedsgrader, der passer til eleverne uanset deres dygtighed. Opgaver med høj statistisk informationsværdi vil ligeledes kunne nedbringe usikkerheden. Ofte vil der være tale om opgaver med flere delopgaver, de såkaldte polytome opgaver.

I testsystemet vises den statistiske usikkerhed på den beregnede elevdygtighed. Læreren har således mulighed for, at vurdere sikkerheden i den enkelte elevs testresultat.

5 Sammenligning mellem tidligere og genberegnede elevdygtigheder

Elevernes dygtighed er genberegnet for de obligatoriske nationale test afholdt i perioden 2015-20. For enkelte test er der også genberegnet elevdygtigheder for frivillige nationale test.

Resultaterne af genberegningen er sammenlignet med de tidligere beregnede elevdygtigheder på elevniveau samt på landsniveau.

5.1 Sammenligning af resultaterne på elevniveau

5.1.1 Sammenligning af elevdygtigheder på logit-skalaen

De genberegnede elevdygtigheder på logit-skalaen kan ikke umiddelbart sammenlignes med de oprindelige elevdygtigheder. Dels har der ikke tidligere været beregnet én fælles elevdygtighed i hver test på den bagvedliggende logit-skala. Dels afhænger parameteriseringen i den enkelte test af den anvendte opgavebank, hvorfor det ikke er muligt at sammenligne niveauet af en elevdygtighed mellem forskellige logit-skalaer. Derimod er det muligt at sammenligne elevernes indbyrdes rangorden i forhold til deres beregnede dygtighed. Hvis elev A havde en højere beregnet dygtighed end elev B i de oprindelige tre profilområder, kan det undersøges, om elev A også har en højere beregnet dygtighed end elev B efter genberegningen.

Til at vurdere, hvorvidt rangordenen mellem elevdygtigheder er bibeholdt efter genberegningen, undersøges korrelationen mellem de tidligere elevdygtigheder og de genberegnede. Korrelationer angiver graden af sammenhæng mellem to variable.

Da der ikke tidligere har været beregnet én fælles elevdygtighed i hver test på den bagvedliggende logit-skala, sammenlignes den genberegnede elevdygtighed på den nye fælles skala derfor med de tidligere beregnede elevdygtigheder i hver af de tre profilområder, som en test består af.

Korrelationskoefficienterne fremgår af tabel 7. De høje korrelationer (0,74 – 0,95) viser, at rangordenen blandt elevers dygtighed er bibeholdt efter samling af de tre profilområder i hver test til én samlet test, og efter at opgavernes sværhedsgrader er genberegnet.

Korrelationerne er beregnet for den samlede periode 2015-2020. Korrelationerne beregnet i hvert skoleår viser en identisk sammenhæng mellem genberegnede og oprindelige dygtigheder (bilag 5 i bilagsnotatet).

Sammenhængen mellem den genberegnede elevdygtighed og den oprindeligt beregnede elevdygtighed er lavest for sprogforståelse i dansk, læsning 8. klasse, hvor korrelationen er på 0,74. Sammenhængen er højest i engelsk, læsning 4. klasse, hvor korrelationen er på 0,95.

Tabel 7. Korrelationen mellem oprindelige elevdygtigheder i hvert profilområde og genberegnede elevdygtigheder på samlet skala

Fag	Klassetrin	Korrelation med genberegnet elevdygtighed for hvert profilområde		
		Profilområde 1	Profilområde 2	Profilområde 3
Dansk, læsning	2. klasse	0,75	0,88	0,91
	4. klasse	0,85	0,87	0,90
	6. klasse	0,82	0,84	0,86
	8. klasse	0,74	0,85	0,82
Matematik	3. klasse	0,89	0,87	0,89
	6. klasse	0,88	0,89	0,90
	8. klasse	0,93	0,93	0,91
Engelsk	4. klasse	0,95	0,93	0,88
	7. klasse	0,90	0,92	0,94
Fysik/kemi	8. klasse	0,83	0,82	0,83
Biologi	8. klasse	0,82	0,83	0,81
Geografi	8. klasse	0,83	0,84	0,84
Dansk som andetsprog	5. klasse	0,88	0,89	0,90
	7. klasse	0,89	0,88	0,89

Note: Pearsons korrelationskoefficient. Elevdygtigheder i perioden 2015-2020.
Kilde: Styrelsen for It og Læring

Sammenhængen mellem den genberegnede elevdygtighed og den oprindeligt beregnede elevdygtighed er ligeledes illustreret i de forskellige scatterplots i bilag 6 i bilagsnotatet.

Generelt for alle scatterplots gælder det, at punkterne er centreret omkring en ret linje, hvilket vidner om en stærk sammenhæng mellem den genberegnede fælles elevdygtighed og den oprindeligt beregnede elevdygtighed i hvert profilområde.

5.1.1.1 Regressionsanalyser

For yderligere at undersøge, hvorvidt de oprindelige elevdygtigheder i hvert af en tests tre profilområder bidrager til at forklare den genberegnede samlede elevdygtighed, er der fortaget lineære regressionsanalyser for alle fag og klassetrin.

Den teoretiske model har den genberegnede samlede elevdygtighed, θ, som den afhængige variabel og de oprindelige elevdygtigheder for hvert profilområde, θ_1, θ_2 *og* θ_3 som forklarende variable. Modellen er således:

$$\theta = \alpha_0 + \alpha_1\theta_1 + \alpha_2\theta_2 + \alpha_3\theta_3 + u_i \,,$$

hvor α_0 er konstantledet og u_i fejlledet.

I bilag 7 i bilagsnotatet findes en samlet tabel med resultatet af regressionsanalyserne. På tværs af alle fag og klassetrin har de oprindelige elevdygtigheder i alle tre profilområder en stærk statistisk signifikant og positiv effekt på de genberegnede fælles elevdygtigheder. De oprindelige elevdygtigheder kan tilsammen forklare mindst 95 procent (R^2 ligger mellem 0,95 og 0,995) af variationen mellem de genberegnede elevdygtigheder i hver test.

Der er således en meget høj grad af overensstemmelse mellem elevdygtighederne beregnet på den nye samlede skala og de oprindelige beregnede elevdygtigheder i de enkelte profilområder. Derudover indikerer de lave estimater på regressionens usikkerhed (bilag 6), at der er en stor præcision i sammenhængen mellem de genberegnede og de oprindelige elevdygtigheder.

5.1.2 Sammenligning af elevdygtigheder på pointskalaen

Elevernes beregnede dygtighed på pointskalaen (0-50) sammenlignes med elevernes oprindelige samlede tilbagemelding på percentilskalaen (1-100). Den enkelte elevs samlede dygtighed på percentilskalaen beregnes som gennemsnit af elevens dygtighed på percentilskalaen i hvert af de tre profilområder i en test.

Korrelationerne mellem de oprindeligt beregnede elevdygtigheder på percentilskalaen og de genberegnede elevdygtigheder omregnet til pointskalaen ligger mellem 0,93 og 0,97 (tabel 8). Der er således en meget stor sammenhæng mellem elevernes oprindelige dygtighed på percentilskalaen og de genberegnede dygtigheder på pointskalaen.

Tabel 8. Korrelation mellem de oprindelige (percentilskala) og de genberegnede elevdygtigheder (pointskala) på elevniveau

	Korrelation					
Fag og klassetrin	**2014/15**	**2015/16**	**2016/17**	**2017/18**	**2018/19**	**2019/20**
Dansk, læsning 2. kl.	0,95	0,95	0,95	0,95	0,95	0,94
Dansk, læsning 4. kl.	0,96	0,96	0,96	0,96	0,96	0,95
Dansk, læsning 6. kl.	0,96	0,97	0,96	0,96	0,96	0,95
Dansk, læsning 8. kl.	0,96	0,96	0,95	0,95	0,95	0,95
Matematik 3. kl.	-	0,96	0,96	0,96	0,96	0,95
Matematik 6. kl.	-	-	-	0,96	0,96	0,95
Matematik 8. kl.	-	-	-	0,97	0,97	0,96
Engelsk 4. kl.	-	-	-	0,97	0,96	0,96
Engelsk 7. kl.	0,96	0,97	0,97	0,97	0,97	0,97
Fysik/kemi 8. kl.	-	0,94	0,94	0,93	0,93	0,94
Biologi 8. kl.	0,96	0,96	0,96	0,96	0,96	-
Geografi 8. kl.	-	0,96	0,96	0,95	0,95	-
Dansk som andetsprog 5. kl.	-	-	0,96	0,96	0,96	-
Dansk som andetsprog 7. kl.	-	-	0,96	0,96	0,96	-

Note: Pearsons korrelationskoefficient mellem tidligere elevdygtigheder og de genberegnede elevdygtigheder i perioden 2015-2020. I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I matematik 6. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2017/18, hvorfor der først er genberegnet herfra. I fysik/kemi og geografi 8. klasse blev opgavebanken gennemgået og opdateret i 2015, hvorfor der først er genberegnet fra og med 2015/16. Testen i matematik 8. klasse og i engelsk 4. klasse er nye test fra og med 2017/18. I dansk som andetsprog, der kun er frivillig test, er genberegningerne først foretaget fra og med efteråret 2016.
Kilde: Styrelsen for It og Læring

Korrelationerne mellem de oprindeligt beregnede elevdygtigheder på percentilskalaen og på pointskalaen er stort set den samme over alle årene med meget få udsving. For alle test i alle årene er sammenhængen mellem de oprindelige og de genberegnede elevdygtigheder statistisk signifikant.

Der er således en meget stor overensstemmelse mellem de oprindeligt beregnede og formidlede elevdygtigheder på percentilskalaen og de genberegnede elevdygtigheder omregnet til pointskalaen.

5.2 Sammenligning af resultaterne på landsniveau

5.2.1 Sammenligning af elevdygtighederne på pointskalaen

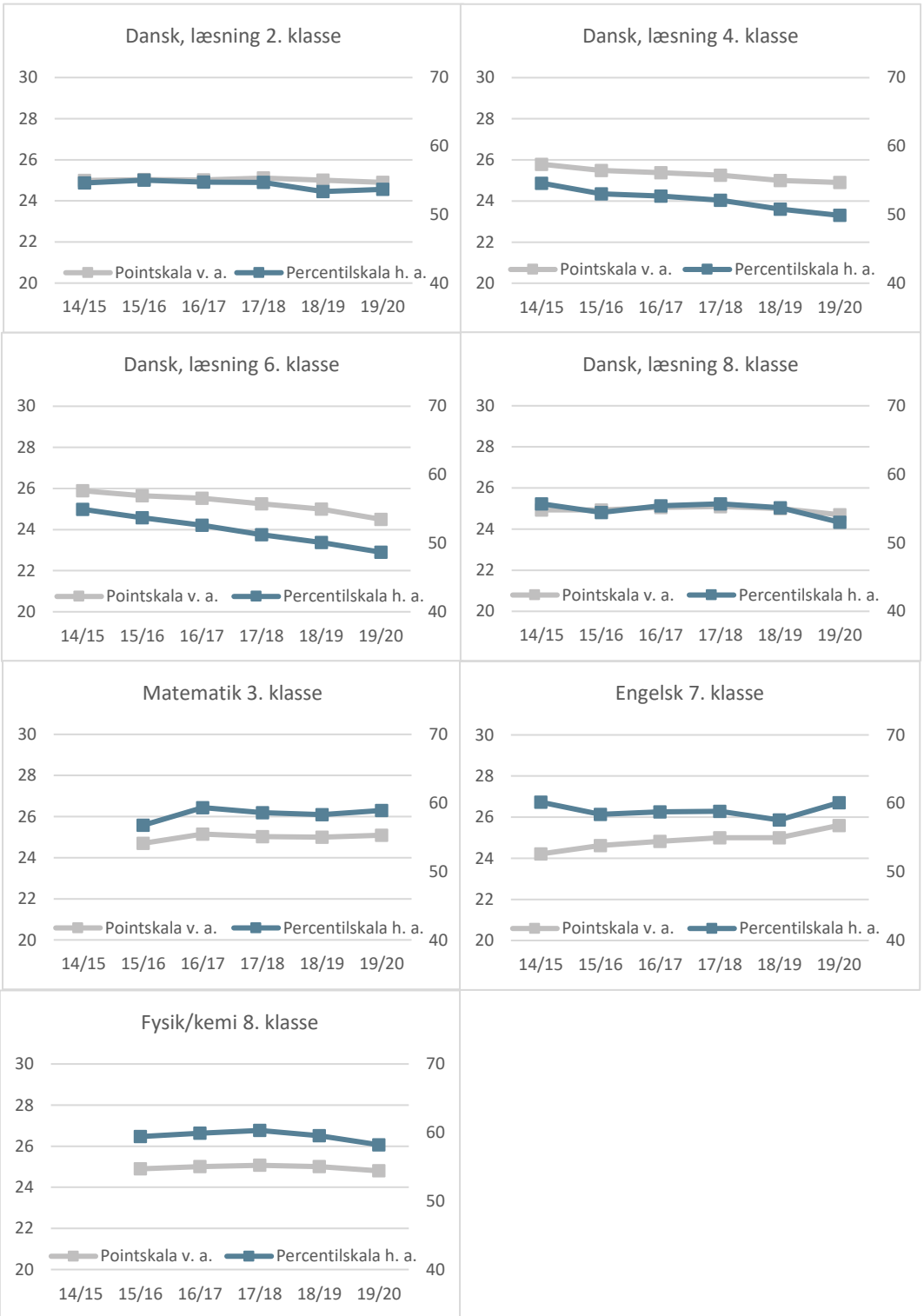
Landsresultaterne opgjort på percentilskalaen kaldes den nationale præstationsprofil. Disse er sammenlignet med landsresultaterne opgjort på pointskalaen efter genberegning af elevdygtighederne (tabel 9).

Tabel 9. Landsresultaterne i de obligatoriske nationale test på henholdsvis percentilskalaen (1-100) og på pointskalaen (0-50)

Fag og klassetrin	Skala	2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
Dansk, læsning 2. kl.	Percentil	55	55	55	55	53	54
	Point	25,0	25,0	25,0	25,1	25,0	24,9
Dansk, læsning 4. kl.	Percentil	55	53	53	52	51	50
	Point	25,8	25,5	25,4	25,3	25,0	24,9
Dansk, læsning 6. kl.	Percentil	55	54	53	51	50	49
	Point	25,9	25,7	25,5	25,2	25,0	24,5
Dansk, læsning 8. kl.	Percentil	56	54	55	56	55	53
	Point	24,9	24,9	25,1	25,1	25,0	24,7
Matematik 3. kl.	Percentil	-	57	59	59	58	59
	Point	-	24,7	25,2	25,0	25,0	25,1
Matematik 6. kl.	Percentil	-	-	-	56	56	55
	Point	-	-	-	25,0	25,0	24,7
Matematik 8. kl.	Percentil	-	-	-	50	50	48
	Point	-	-	-	25,0	25,0	24,6
Engelsk 4. kl.	Percentil	-	-	-	50	52	57
	Point	-	-	-	24,7	25,0	26,0
Engelsk 7. kl.	Percentil	60	58	59	59	58	60
	Point	24,2	24,6	24,8	25,0	25,0	25,6
Fysik/kemi 8. kl.	Percentil	-	60	60	60	60	58
	Point	-	24,9	25,0	25,1	25,0	24,8
Biologi 8. kl.	Percentil	54	60	59	-	-	-
	Point	24,6	24,8	24,7	-	-	-
Geografi 8. kl.	Percentil	-	61	62	-	-	-
	Point	-	24,5	24,6	-	-	-

Note: I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I matematik 6. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2017/18, hvorfor der først er genberegnet herfra. I fysik/kemi og geografi 8. klasse blev opgavebanken gennemgået og opdateret i 2015, hvorfor der først er genberegnet fra og med 2015/16. Testen i matematik 8. klasse og i engelsk 4. klasse er nye test fra og med 2017/18. Fra og med 2017/18 er testen i biologi og geografi ikke obligatorisk. I 2019/20 indgår kun resultater fra de obligatoriske nationale test, der blev gennemført som et repræsentativt udsnit af landets skoler. **Kilde:** Styrelsen for It og Læring

Figur 3. Udviklingen i landsresultater på henholdsvis pointskalaen (venstre akse) og percentilskalaen (højre akse)



Note: I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I fysik/kemi 8. klasse blev opgavebanken gennemgået og opdateret i 2015, hvorfor der først er genberegnet fra og med 2015/16.
Kilde: Styrelsen for It og Læring

Udviklingen er ligeledes illustreret i figur 3 for de obligatoriske test i dansk, læsning i 2., 4., 6. og 8. klasse, matematik i 3. klasse, engelsk i 7. klasse og fysik/kemi i 8. klasse.

Udviklingen i landsresultaterne i de obligatoriske nationale test i dansk, læsning, i matematik og i fysik/kemi er næsten den samme på den oprindelige percentilskala og på den nye pointskala baseret på genberegnede elevdygtigheder (figur 3). Da de to skalaer er forskellige, er det udelukkende udviklingstendensen, der kan sammenlignes.

I engelsk 7. klasse viser de genberegnede elevdygtigheder opgjort på pointskalaen på landsplan en lille stigning over årene. De oprindelige resultater på percentilskalaen viste et lille fald i 2015/16, hvorefter landsresultatet næsten har været konstant over årene[17]. I 2019/20 er der igen tale om den samme udviklingstendens på de to skalaer.

Med undtagelse af engelsk 7. klasse er udviklingen i landsresultaterne således sammenfaldende mellem den oprindelige percentilskala og pointskalaen.

5.2.2 Sammenligning af elevdygtighederne på den kriteriebaserede skala

Den kriteriebaserede skala anvendes alene til formidling af testresultaterne i dansk, læsning og matematik.

På landsplan er andelen af elever, der er gode til dansk, læsning og matematik, stort set uændret efter genberegning af elevernes testresultater (tabel 10).

I skoleåret 2018/19 er der fuld overensstemmelse mellem den oprindelige opgørelse af andelen af elever med mindst et godt testresultat og den nye baseret på genberegnede elevdygtigheder.

I dansk, læsning 2. klasse er andelen af elever, der har opnået et godt testresultat, 1 procentpoint lavere efter genberegningen i skoleårene 2015/16 og 2016/17 og 2 procentpoint lavere i skoleåret 2019/20.

I dansk, læsning 4. klasse er andelen af elever, der har opnået et godt testresultat, 2 procentpoint højere efter genberegningen i skoleåret 2014/15 og 1 procentpoint højere i 2015/16.

I dansk, læsning 6. klasse er andelen af elever, der har opnået et godt testresultat, 1 procentpoint højere efter genberegningen i skoleåret 2014/15 og 2 procentpoint lavere i skoleåret 2019/20, og i dansk, læsning 8. klasse er andelen steget med 1 procentpoint i 2014/15 og 2015/16 efter genberegningen.

I matematik 3. klasse er andelen af elever, der har opnået et godt testresultat, 1 procentpoint lavere efter genberegningen i skoleårene 2016/17 og 2017/18, og i 6. klasse er andelen 1 procentpoint højere i skoleåret 2019/20. I matematik 8. klasse er der ingen ændring på landsplan efter genberegningen.

[17] I to af profilområderne er der tale om et fald i landsresultaterne over årene, mens der er en stigning i landsresultaterne i det tredje profilområde på den oprindelige percentilskala.

Tabel 10. Andel af elever, der er gode til læsning og matematik i den oprindelige opgørelse (Før) og i opgørelsen baseret på genberegnede elevdygtigheder (Nu).

Fag og klassetrin		2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
Dansk, læsning 2. kl.	Før	75 %	76 %	76 %	76 %	73 %	74 %
	Nu	75 %	75 %	75 %	76 %	73 %	72 %
	Forskel	**0**	**÷1**	**÷1**	**0**	**0**	**÷2**
Dansk, læsning 4. kl.	Før	69 %	67 %	67 %	66 %	64 %	62 %
	Nu	71 %	68 %	67 %	66 %	64 %	62 %
	Forskel	**+2**	**+1**	**0**	**0**	**0**	**0**
Dansk, læsning 6. kl.	Før	72 %	71 %	70 %	68 %	66 %	64 %
	Nu	73 %	71 %	70 %	68 %	66 %	62 %
	Forskel	**+1**	**0**	**0**	**0**	**0**	**÷2**
Dansk, læsning 8. kl.	Før	77 %	76 %	78 %	78 %	77 %	74 %
	Nu	78 %	77 %	78 %	78 %	77 %	74%
	Forskel	**+1**	**+1**	**0**	**0**	**0**	**0**
Matematik 3. kl.	Før	-	73 %	77 %	76 %	75 %	76 %
	Nu	-	73 %	76 %	75 %	75 %	76 %
	Forskel	-	**0**	**÷1**	**÷1**	**0**	**0**
Matematik 6. kl.	Før	-	-	-	77 %	77 %	74 %
	Nu	-	-	-	77 %	77 %	75 %
	Forskel	-	-	-	**0**	**0**	**+1**
Matematik 8. kl.	Før	-	-	-	79 %	79 %	76 %
	Nu	-	-	-	79 %	79 %	76 %
	Forskel	-	-	-	**0**	**0**	**0**

Note: Andel gode dækker over andel med en god, en rigtig god eller en fremragende præstation. I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I matematik 6. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2017/18, hvorfor der først er genberegnet herfra. Testen i matematik 8. klasse er ny test fra og med 2017/18. I 2019/20 indgår kun resultater fra de obligatoriske nationale test, der blev gennemført som et repræsentativt udsnit af landets skoler.
Kilde: Styrelsen for It og Læring

Efter genberegningen af elevernes dygtigheder er der ikke ændret på den oprindelige udvikling i elevernes testresultater. Der er et generelt fald i andelen af elever med et godt testresultat i dansk, læsning i 4. og 6. klasse over årene, mens der i 2. og 8. klasse er tale om år til år variationer.

På landsplan er andelen af elever, der er blandt de dygtigste til dansk, læsning og matematik, stort set uændret efter genberegning af elevernes testresultater (tabel 11).

I skoleåret 2018/19 er der fuld overensstemmelse mellem den oprindelige opgørelse og den nye baseret på genberegnede elevdygtigheder.

Tabel 11. Andel af elever, der er blandt de dygtigste til læsning og matematik, i den oprindelige opgørelse (Før) og i opgørelsen baseret på genberegnede elevdygtigheder (Nu).

Fag og klassetrin		2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
Dansk, læsning 2. kl.	Før	7 %	9 %	9 %	9 %	9 %	10 %
	Nu	8 %	8 %	8 %	9 %	9 %	10 %
	Forskel	**+1**	**÷1**	**÷1**	**0**	**0**	**0**
Dansk, læsning 4. kl.	Før	9 %	8 %	9 %	8 %	8 %	6 %
	Nu	10 %	9 %	9 %	8 %	8 %	8 %
	Forskel	**+1**	**+1**	**0**	**0**	**0**	**+2**
Dansk, læsning 6. kl.	Før	6 %	6 %	5 %	5 %	4 %	4 %
	Nu	6 %	6 %	5 %	5 %	4 %	4 %
	Forskel	**0**	**0**	**0**	**0**	**0**	**0**
Dansk, læsning 8. kl.	Før	12 %	12 %	14 %	14 %	14 %	14 %
	Nu	13 %	14 %	14 %	14 %	14 %	14 %
	Forskel	**+1**	**+2**	**0**	**0**	**0**	**0**
Matematik 3. kl.	Før	-	9 %	11 %	10 %	11 %	11 %
	Nu	-	9 %	11 %	11 %	11 %	11 %
	Forskel	-	**0**	**0**	**+1**	**0**	**0**
Matematik 6. kl.	Før	-	-	-	8 %	9 %	10%
	Nu	-	-	-	8 %	9 %	8 %
	Forskel	-	-	-	**0**	**0**	**÷2**
Matematik 8. kl.	Før	-	-	-	6 %	7 %	5 %
	Nu	-	-	-	6 %	7 %	5 %
	Forskel	-	-	-	**0**	**0**	**0**

Note: Andel dygtigste dækker over andel med en fremragende præstation. I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I matematik 6. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2017/18, hvorfor der først er genberegnet herfra. Testen i matematik 8. klasse er ny test fra og med 2017/18. I 2019/20 indgår kun resultater fra de obligatoriske nationale test, der blev gennemført som et repræsentativt udsnit af landets skoler.
Kilde: Styrelsen for It og Læring

Helt tilsvarende er andelen af elever, der er blandt de dårligste til dansk, læsning og matematik på landsplan, stort set uændret efter genberegning af elevernes testresultater (tabel 12).

I skoleåret 2018/19 er der fuld overensstemmelse mellem den oprindelige opgørelse og den nye baseret på genberegnede elevdygtigheder.

Tabel 12. Andel af elever, der er blandt de dårligste til læsning og matematik, i den oprindelige opgørelse (Før) og i opgørelsen baseret på genberegnede elevdygtigheder (Nu).

Fag og klassetrin		2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
Dansk, læsning 2. kl.	Før	10 %	9 %	9 %	8 %	9 %	10 %
	Nu	9 %	9 %	9 %	8 %	9 %	10 %
	Forskel	**÷1**	**0**	**0**	**0**	**0**	**0**
Dansk, læsning 4. kl.	Før	13 %	13 %	14 %	14 %	15 %	15 %
	Nu	11 %	13 %	14 %	14 %	15 %	16 %
	Forskel	**÷2**	**0**	**0**	**0**	**0**	**+1**
Dansk, læsning 6. kl.	Før	11 %	11 %	11 %	12 %	13 %	15 %
	Nu	10 %	10 %	11 %	12 %	13 %	15 %
	Forskel	**÷1**	**÷1**	**0**	**0**	**0**	**0**
Dansk, læsning 8. kl.	Før	10 %	9 %	9 %	9 %	9 %	12 %
	Nu	9 %	9 %	9 %	8 %	9 %	12 %
	Forskel	**÷1**	**0**	**0**	**÷1**	**0**	**0**
Matematik 3. kl.	Før	-	12 %	10 %	11 %	12 %	12 %
	Nu	-	12 %	11 %	11 %	12 %	11 %
	Forskel	-	**0**	**+1**	**0**	**0**	**+1**
Matematik 6. kl.	Før	-	-	-	10 %	10 %	12 %
	Nu	-	-	-	10 %	10 %	12 %
	Forskel	-	-	-	**0**	**0**	**0**
Matematik 8. kl.	Før	-	-	-	6 %	6 %	8 %
	Nu	-	-	-	6 %	6 %	8 %
	Forskel	-	-	-	**0**	**0**	**0**

Note: Andel dårligste dækker over andel med en mangelfuld eller ikke tilfredsstillende præstation. I matematik 3. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2015/16, hvorfor der først er genberegnet herfra. I matematik 6. klasse er profilområdet 'Matematik i anvendelse' erstattet med 'Statistik og sandsynlighed' fra og med 2017/18, hvorfor der først er genberegnet herfra. Testen i matematik 8. klasse er ny test fra og med 2017/18. I 2019/20 indgår kun resultater fra de obligatoriske nationale test, der blev gennemført som et repræsentativt udsnit af landets skoler.
Kilde: Styrelsen for It og Læring

6 Testenes reliabilitet og kriterievaliditet

Testenes reliabilitet og validitet har tidligere været beskrevet, senest i forbindelse med evalueringen af de nationale test[18].

I dette afsnit er reliabiliteten og kriterievaliditeten beregnet igen. Denne gang ved anvendelse af de genberegnede elevdygtigheder.

6.1 Testenes reliabilitet

Testens evne til at nå det samme resultat ved gentagne målinger kaldes testens reliabilitet. Reliabiliteten kan også opfattes som et udtryk for testens evne til at adskille elever med forskellig dygtighed på korrekt vis.

Reliabiliteten afhænger af såvel den statistiske usikkerhed på elevdygtigheden, SEM, samt af spredningen mellem elevernes dygtighed i den pågældende test. Des mindre SEM, desto højere reliabilitet. Samtidig giver en lille spredning mellem elevernes dygtighed en lavere reliabilitet. Hvis alle elever har samme dygtighed, er spredningen og reliabiliteten lig nul. En stor spredning i elevernes dygtighed vil give en højere reliabilitet, der nærmer sig 1.

I RUMM[19] beregnes et indeks for reliabiliteten: *Person Separation Index*'[20].

Person separation indeks (PSI) beregnes som:

$$r_{\theta\theta} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_\theta^2},$$

hvor θ er elevens estimerede dygtighed, og σ_θ^2 beregnes som variansen mellem elevernes dygtigheder blandt de elever, der gennemfører en test. σ_ε^2 er usikkerheden på den enkelte elevs beregnede dygtighed.

Reliabiliteten, efter genberegning af elevernes dygtighed, ligger i intervallet 0,85-0,95 (tabel 13).

Der findes forskellige anbefalinger til niveauet af reliabilitet. I Streiner[21] anføres, at en optimal reliabilitet ikke bør være under 0,70. I Nunnally & Bernstein[22] anføres, at reliabiliteten for test, der anvendes til vurderinger på individniveau, ikke bør være under 0,90. En reliabilitet på 0,90 svarer til en SEM på 0,30, hvis spredningen mellem elevernes dygtighed er lig 1.

[18] https://www.uvm.dk/folkeskolen/elevplaner-nationale-test--trivselsmaaling-og-sprogproever/nationale-test/evaluering-af-de-nationale-test

[19] RUMM Laboratory Pty Ltd.

[20] Christensen K.B., Kreiner S., Mesbah M. (editors): Rasch Models in Health. John Wiley & Sons, Inc., USA. 2013

[21] Streiner, D. L., Norman G. R. (1995): Health Measurement Scales – A Practical Guide to Their Development and Use. Oxford University Press

[22] Nunnally, J. C., Bernstein, I. H. (1994): Psychometric Theory. New York. McGraw-Hill.

I alle test ligger reliabiliteten over 0,85, og i 8 ud af de 14 test ligger reliabiliteten over 0,90. Reliabiliteten er lavest i dansk, læsning 8. klasse samt i fysik/kemi, biologi og geografi i 8. klasse. Dette hænger sammen med, at spredningen mellem elevernes dygtighed i disse test er en del mindre end i de øvrige test.

Tabel 13. Reliabiliteten af de nationale test

Fag	Klassetrin	Reliabilitet
Dansk, læsning	2. klasse	0,89
	4. klasse	0,90
	6. klasse	0,88
	8. klasse	0,85
Matematik	3. klasse	0,91
	6. klasse	0,92
	8. klasse	0,94
Engelsk	4. klasse	0,95
	7. klasse	0,94
Fysik/kemi	8. klasse	0,85
Biologi	8. klasse	0,85
Geografi	8. klasse	0,86
Dansk som andetsprog	5. klasse	0,92
	7. klasse	0,92

Note: Person Separation Index
Kilde: Styrelsen for It og Læring

Før sammenlægningen af de tre oprindelige profilområder til beregning af én samlet elevdygtighed i hver test lå reliabiliteten i de enkelte profilområder i intervallet 0,66 – 0,91.

Den øgede statistiske sikkerhed på elevernes dygtighed har således forbedret testenes reliabilitet væsentligt.

6.2 Testenes kriterievaliditet

For at få en indikation af om testene måler det samme som andre tilsvarende test og prøver, undersøges sammenhængen mellem elevernes testresultat i de nationale test og deres efterfølgende præstation i de relevante dele af folkeskolens prøver i 9. klasse.

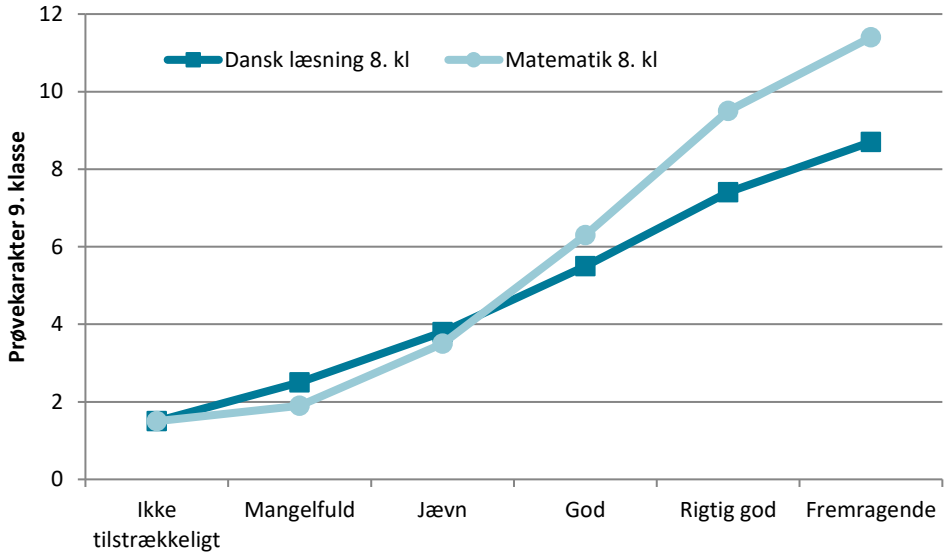
Elevernes karakter i dansk, læsning i folkeskolens prøve i 9. klasse i 2018/19 er sammenholdt med elevernes testresultater i de obligatoriske nationale test i dansk, læsning i 8. klasse i 2017/18. Tilsvarende er elevernes karakter i matematik uden hjælpemidler i folkeskolens prøve i 9. klasse i 2018/19 sammenholdt med elevernes testresultater i de obligatoriske nationale test i matematik i 8. klasse i 2017/18 (figur 4).

I både dansk, læsning og matematik er der en statistisk signifikant sammenhæng mellem elevernes testresultat i de nationale test i 8. klasse og deres karakter fra tilsvarende fag i folkeskolens prøver i 9. klasse.

Gruppen af elever, der opnår en jævn præstation som resultat i de obligatoriske nationale test i dansk, læsning 8. klasse, får i gennemsnit 3,8 i karakter ved folkeskolens prøver i 9.

klasse, mens gruppen af elever, der opnår en rigtig god præstation i de obligatoriske nationale test i dansk, læsning 8. klasse, får 7,4 i gennemsnit i karakter ved folkeskolens prøver i 9. klasse.

Figur 4. Sammenhængen mellem resultaterne fra de nationale test og karakterer fra folkeskolens prøver i 9. klasse i 2018/19



Kilde: Styrelsen for It og Læring

Den samme tydelige sammenhæng ses mellem elevernes testresultater i matematik i 8. klasse og deres karakter året efter ved folkeskolens prøve i matematik uden hjælpemidler i 9. klasse.

Elevernes fordeling på karaktererne ved folkeskolens prøver i 9. klasse i forhold til elevernes testresultater på den kriteriebaserede skala i de nationale test ses i tabel 14.

Blandt de elever, der opnår en ikke tilstrækkelig præstation i dansk, læsning i de nationale test i 8. klasse, får 58 procent højst karakteren 0 i dansk, læsning året efter ved folkeskolens prøve i 9. klasse.

Blandt de elever, der opnår en rigtig god præstation i dansk, læsning i de nationale test i 8. klasse, får 75 procent karakteren 7 eller 10 i dansk, læsning året efter ved folkeskolens prøve i 9. klasse.

Blandt de elever, der opnår en god præstation i matematik i de nationale test i 8. klasse, får 84 procent karakteren 4 eller 7 i matematik uden hjælpemidler året efter ved folkeskolens prøve i 9. klasse.

Blandt de elever, der opnår en fremragende præstation i matematik i de nationale test i 8. klasse, får 98 procent karakteren 10 eller 12 i matematik uden hjælpemidler året efter ved folkeskolens prøve i 9. klasse.

Tabel 14. Elevernes testresultat i de nationale test sammenholdt med karakteren fra folkeskolens prøve i 9. klasse i 2018/19. Andel elever (procent)

Fag og klassetrin	Testresultat	Karakter (9. kl.)							I alt
		-3	0	2	4	7	10	12	
Dansk læsning 8. klasse	Ikke tilstrækkelig	0	58	21	13	7	1	0	100
	Mangelfuld	0	31	30	27	11	1	0	100
	Jævn	0	12	24	39	23	2	0	100
	God	0	4	11	32	43	9	1	100
	Rigtig god	0	1	4	16	47	28	5	100
	Fremragende	0	1	2	8	31	39	20	100
Matematik 8. klasse	Ikke tilstrækkelig	0	50	31	13	5	1	0	100
	Mangelfuld	0	32	43	22	3	1	0	100
	Jævn	0	7	30	52	11	0	0	100
	God	0	0	3	28	56	11	1	100
	Rigtig god	0	0	0	2	27	51	20	100
	Fremragende	0	0	0	0	2	24	74	100

Kilde: Styrelsen for It og Læring

Der er således fortsat en tydelig sammenhæng mellem elevernes resultater i de nationale test i dansk, læsning og i matematik i 8. klasse og deres karakterer i folkeskolernes prøve året efter i 9. klasse.