BØRNE- OG
UNDERVISNINGS-
MINISTERIET
STYRELSEN
FOR IT OG LÆRING

Sammenfatning

Evaluering af de statistiske aspekter ved de nationale test

Opgavebeskrivelse

Evalueringen af de nationale test består af to dele:

1. Validering af teknisk beregning
2. Undersøgelse af betydningen og brugen af de nationale test

Under delopgave 1 skal følgende evalueringsspørgsmål besvares:

1. *Regner de nationale test rigtigt?*
   Ved besvarelse af spørgsmålet skal den kritik der rejses af den statistiske sikkerhed og reliabiliteten i de nationale test adresseres. Herunder skal det klarlægges om:
   a. opgavernes sværhedsgrader stadig er korrekte?
   b. opgaverne fortsat passer til Rasch-modellen?
   c. det er muligt at forbedre den adaptive algoritme med henblik på at reducere den statistiske usikkerhed?

2. Det skal afdækkes, om sikkerheden i målingerne af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder? Herunder sigter spørgsmålet på at klarlægge følgende:
   a. Kan det påvises, at profilområderne måler forskellige aspekter af den samme bagvedliggende færdighed?
   b. Som følge af spørgsmål a: Kan testresultaterne fra profilområderne slås sammen og dermed forbedre sikkerheden i testene?

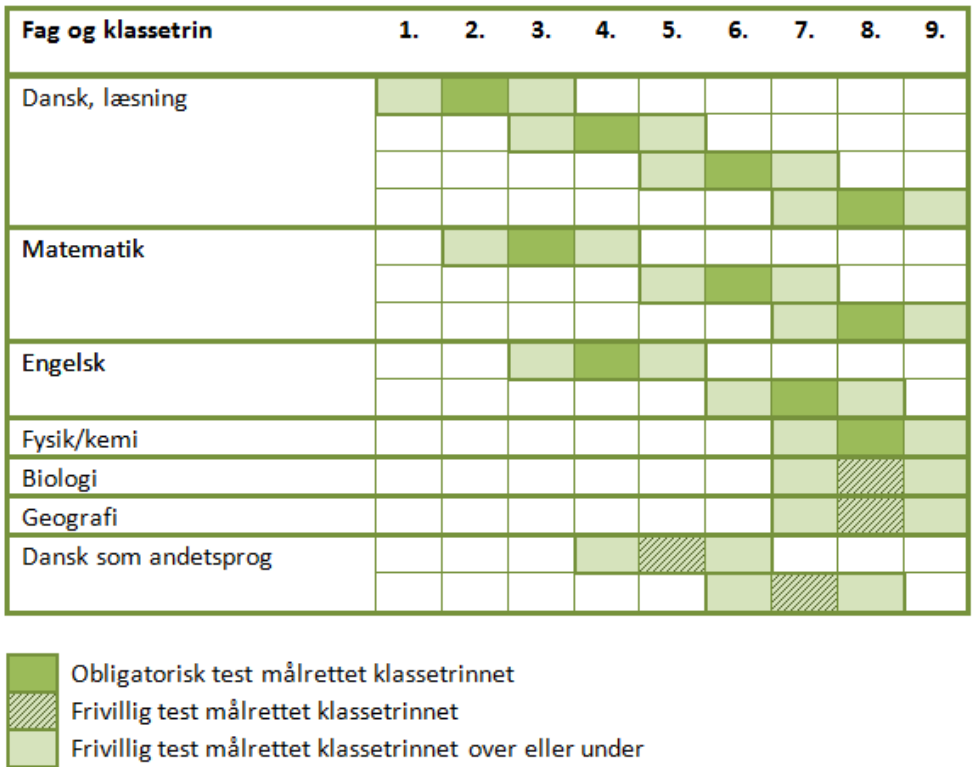
Indledning

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

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.

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

Figur 1 Frivillige og obligatoriske nationale test



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

Notatet indeholder et kort resume af de gennemførte analyser, der vedrører delopgave 1. Børne- og undervisningsministeriet (BUVM) har tidligere undersøgt mange af evalueringens temaer og formidlet disse på www.uvm.dk[2]. Evalueringen af de statistiske aspekter ved de nationale test samler de tidligere gennemførte analyser og supplerer disse med opdaterede data og nye analyser. Notatet indeholder følgende afsnit:

- Fungerer algoritmen korrekt og vælges de rigtige opgaver i det adaptive forløb
- Måler testene det samme som andre tilsvarende test og prøver
- Den statistiske usikkerhed og testenes reliabilitet
- Er opgavernes sværhedsgrad korrekt bestemt
- Kan elevernes beregnede dygtighed fra tre profilområder samles til én vurdering af dygtigheden

I de enkelte afsnit er der henvisning til de bagvedliggende mere udførlige notater. Disse notater er samlet i rapporten *Evaluering af de statistiske aspekter ved de nationale test*.

[2] https://www.uvm.dk/folkeskolen/elevplaner-nationale-test--trivselsmaaling-og-sprogproever/nationale-test/om-de-nationale-test

Fungerer algoritmen i test- og prøvesystemet korrekt og vælges de rigtige opgaver i det adaptive forløb

I evalueringen af de nationale test skal følgende spørgsmål besvares:

1. Regner de nationale test rigtigt?

For at svare på spørgsmålet er det først og fremmest vigtigt at vurdere, om algoritmen i testsystemet fungerer efter hensigten.

I testsystemets adaptive algoritme vælges opgaverne således, at de bedst muligt passer til elevens dygtighed. Efter hver besvarelse beregnes elevens dygtighed og den næste opgave vælges. Først søges i opgavebanken i et lille interval omkring den sværhedsgrad, der passer til elevens dygtighed. Findes ingen opgaver i dette interval, da udvides intervallet indtil, der findes en passende opgave.

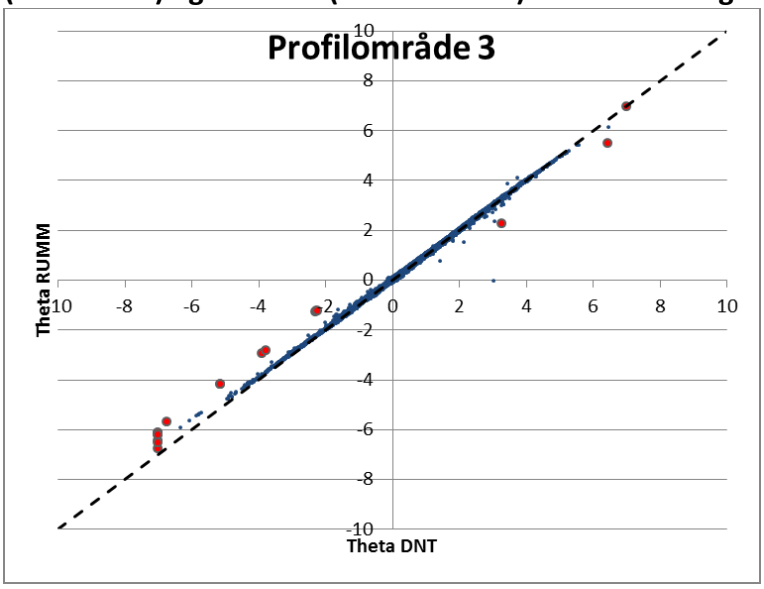
I materialet til evalueringen er medtaget eksempler fra elevers testforløb. Eksemplerne viser, at algoritmen vælger opgaverne som forudsat.

For yderligere at tjekke om testsystemet beregner elevernes dygtighed og den statistiske usikkerhed korrekt, er elevdygtighederne og usikkerheden kontrolberegnet i et kommercielt softwareprogram. Kontrolberegningerne er foretaget i softwareprogrammet RUMM[3], der er udviklet på University of Western Australia, Perth.

Beregningerne viser fuld overensstemmelse mellem de beregnede elevdygtigheder i testsystemet og i RUMM for over 99 procent af forløbene. Figur 2 viser resultaterne fra ét af profilområderne.

[3] www.rummlab.com.au

Figur 2 Sammenhæng mellem beregnet elevdygtighed i testsystemet (Theta DNT) og i RUMM (Theta RUMM). Dansk læsning 8. klasse



Note: Røde dots er elever med ekstreme besvarelser

Kilde: Styrelsen for It og Læring

Den lille andel elevforløb, hvor der er en afvigelse i den beregnede elevdygtighed, er de såkaldte 'ekstreme' forløb, hvor eleven enten har svaret rigtigt på alle opgaver eller forkert på alle opgaver. Beregning af elevdygtighed i disse forløb håndteres en anelse forskelligt i forskellige programmer.

Den adaptive algoritme i testsystemet fungerer ifølge forskrifterne, og elevernes dygtighed og usikkerheden på den beregnede dygtighed beregnes korrekt.

Analyserne er uddybet i:

- Notat 1. Algoritmen i testsystemet og beregning af elevdygtigheden
  - Bilag 1.1. Anvendte skalaer til præsentation af elevernes beregnede dygtigheder
  - Bilag 1.2. Opgavebanken i dansk læsning 8. klasse - sprogforståelse

Måler testene det samme som andre tilsvarende test og prøver

Et andet element i vurderingen af spørgsmålet:
1. Regner de nationale test rigtigt?

er, at undersøge om elevernes resultater fra de nationale test stemmer overens med elevernes resultater fra andre tilsvarende test og prøver. En sådan egenskab omtales som testenes kriterievaliditet.

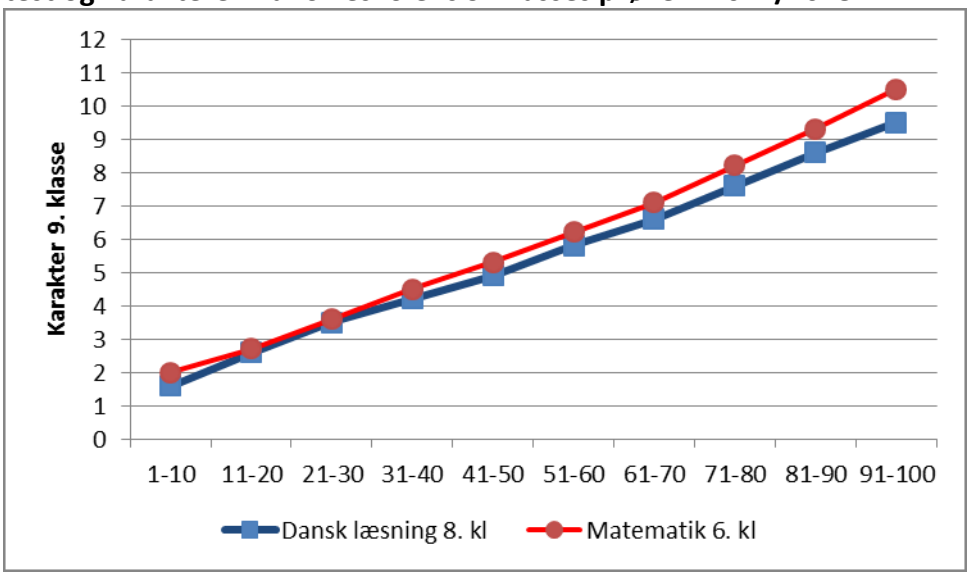
Hvis de nationale test beregner elevernes dygtighed forkert, må det forventes, at sammenhængen mellem elevernes beregnede dygtigheder i de nationale test og resultaterne fra andre test og elevvurderinger er begrænset.

For at få en indikation af om testene samlet set ser ud til at måle det samme som andre tilsvarende test og prøver, kan man se på sammenhængen mellem elevernes testresultat i de nationale test og deres efterfølgende præstation i de relevante dele af standpunktsprøverne i 8. klasse samt i folkeskolens prøver i 9. klasse. Endvidere er det muligt at se på sammenhængen mellem elevernes testresultater i de nationale test og elevernes senere PISA resultater. Begge dele er tidligere undersøgt af BUVM[4,5].

I dansk læsning sammenholdes elevernes beregnede dygtighed i de nationale test i 8. klasse i skoleåret 2016/2017 med de samme elevers karakter fra folkeskolens prøve i læsning i 9. klasse i skoleåret 2017/2018.

Gruppen af elever, der opnår mellem 31 og 40 point[6] i samlet vurdering i de obligatoriske nationale test i dansk, læsning 8. klasse får i gennemsnit 4,2 i karakter ved folkeskolens prøver i dansk læsning i 9. klasse, mens gruppen af elever, der opnår mellem 81 og 90 point i gennemsnit får 8,6 i karakter ved prøven i dansk læsning i 9. klasse (Figur 3).

Figur 3 Sammenhængen mellem resultaterne fra de nationale obligatoriske test og karakterer fra folkeskolens 9. klasses prøver i 2017/2018



Kilde: Styrelsen for It og Læring

[4] https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf16/sep/160912-notat-om-nationale-tests-maaleegenskaber.pdf

[5] https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf17/jan/170110-kriteriebaserede-test-delrapport-1.pdf

[6] I formidlingen til elever og forældre bliver resultaterne på 100-skalaen omsat til: En del under gennemsnittet (1-10), under gennemsnittet (11-35), gennemsnittet (35-65), over gennemsnittet (66-90), en del over gennemsnittet (91-100)

I matematik sammenholdes elevernes beregnede dygtighed i de nationale test i 6. klasse i skoleåret 2014/2015 med de samme elevers karakter fra folkeskolens prøve i matematik uden hjælpemidler i 9. klasse tre år senere i skoleåret 2017/2018.

Gruppen af elever, der opnår mellem 21 og 30 point i samlet vurdering i de obligatoriske nationale test i matematik 6. klasse får i gennemsnit 3,6 i prøvekarakter i matematik uden hjælpemidler tre år senere i 9. klasse, mens gruppen af elever, der opnår mellem 81 og 90 point i gennemsnit får 9,3 i prøvekarakter i matematik i 9. klasse (Figur 3).

Tilsvarende entydige sammenhæng findes mellem testresultater og elevernes standpunktskarakterer i 8. klasse.

Der er naturligvis spredning i resultaterne, dvs. der er elever, der opnår et godt testresultat og efterfølgende en mindre god prøvekarakter og omvendt. Men *gruppen af elever*, der klarer testen med et resultat en del under gennemsnittet, vil også i gennemsnit få en prøvekarakter, der ligger relativt lavt – selv i matematik, hvor der er tre år mellem testafvikling og folkeskolens prøver.

Den samme prædiktive validitet er flere forskere kommet frem til, blandt andet Louise Beuchert & Anne Nandrup fra Aarhus Universitet[7].

I en rapport fra konsulentfirmaet DAMVAD udarbejdet i samarbejde med Svend Kreiner i 2014[8] påvises det endvidere, at der er en tydelig sammenhæng mellem de resultater, eleverne opnår i de nationale test og i den internationale PISA-undersøgelse, jf. boks 1. Dette gælder både for testene i dansk og matematik.

[7] Louise V. Beuchert & Anne B. Nandrup. The Danish National Tests at a Glance. Nationaløkonomisk Tidsskrift 2018:2

[8] PISA-relatering af de kriteriebaserede nationale test. DAMVAD 2014 (https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf17/jan/170110-kriteriebaserede-test-delrapport-1.pdf)

Boks 1. Uddrag af Damvad-rapport om PISA og de nationale test (s. 5):
"Der er en tydelig sammenhæng mellem resultaterne fra de nationale test og resultaterne fra PISA-undersøgelserne. Sammenhængen kan observeres på tværs af profilområder i både læsning og matematik, men er ikke nødvendigvis jævnt fordelt."

"Den tydelige sammenhæng mellem resultaterne fra de nationale test og PISA betyder samtidig, at de to test uafhængigt af hinanden når til relativt enslydende vurderinger af elevers faglige niveauer. Det er en bekræftelse af, at de nationale test siger noget relevant om elevernes faglige niveau i de områder, hvori de testes."

Den faglige færdighed i læsning og matematik, der måles i de nationale test, kan således relateres til såvel udvalgte dele af folkeskolens prøver samt til den internationale PISA-undersøgelse.
Analyserne er uddybet i:

- Notat 2. De nationale tests måleegenskaber
    - Bilag 2.1. Sammenhæng mellem testresultater og karakterer

Den statistiske usikkerhed og testenes reliabilitet
I evalueringen af de nationale test skal følgende kritik belyses:
1. Den kritik der rejses af den statistiske sikkerhed og reliabiliteten skal adresseres.
1c. Herunder skal det klarlægges om det er muligt at forbedre den adaptive algoritme med henblik på at reducere den statistiske usikkerhed.

Den statistiske usikkerhed
Først og fremmest skal det bemærkes, at anvendelsen af en statistisk model, Rasch-modellen[9,10], medfører, at den statistiske usikkerhed på elevernes beregnede dygtighed kan beregnes i de nationale test. Denne usikkerhed bliver formidlet til lærerne via resultatvisningerne i testsystemet. Denne funktionalitet er unik for de nationale test, mens den statistiske usikkerhed på f.eks. elevernes standpunkts- og prøvekarakterer ikke beregnes og formidles.

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

[10] Karl Bang Christensen, Svend Kreiner, Mounir Mesbah (edt): Rasch Models in Health. Wiley 2013.

Den gennemsnitlige statistiske usikkerhed[11] på den beregnede elevdygtighed i de nationale obligatoriske test i 2017/2018 er på 0,46 logit[12].

Hver test gennemføres på 45 minutter og hver test består af test i tre faglige profilområder. Der er således ca. 15 minutter til en test i et profilområde. I gennemsnit når eleverne at besvare 19 opgaver i hvert profilområde på den afsatte tid. De fleste opgaver kan besvares rigtigt eller forkert. Disse kaldes dikotome[13] opgaver. Der findes også opgaver med flere delopgaver (polytome[14] opgaver), hvor eleven kan få 0, 1, 2 eller flere rigtige. Tælles alle delopgaver med, da kan eleverne i gennemsnit nå at besvare 23 opgaver/delopgaver i hvert profilområde.

For at vurdere om en statistisk usikkerhed på 0,46 logit er stor eller lille kan anvendes, at usikkerheden i adaptive test med dikotome opgaver ikke kan blive mindre end $2/\sqrt{n}$, hvor n er antallet af opgaver.

Med 19 dikotome opgaver er den mindst mulige usikkerhed på 0,46, mens der med 23 dikotome opgaver ikke kan opnås en usikkerhed på mindre end 0,42 (Tabel 1).

Tabel 1 Sammenhæng mellem antal opgaver og mindst mulige SEM

Antal opgaver	SEM[1)]		Antal opgaver	SEM[1)]
15	0,52		22	0,43
16	0,50		23	0,42
17	0,49		24	0,41
18	0,47		25	0,40
19	0,46		30	0,37
20	0,45		40	0,32
21	0,44		45	0,30

1) Den statistiske usikkerhed betegnes SEM

Kilde: Styrelsen for It og Læring

Den gennemsnitlige statistiske usikkerhed i de nationale test på 0,46 er således ikke langt fra det mest optimale på 0,42.

[11] Den statistiske usikkerhed på elevernes beregnede dygtighed betegnes SEM

[12] Målingerne af elevdygtigheden og den statistiske usikkerhed foregår på en såkaldt logit-skala. Logits er en transformation med den naturlige logaritme af odds, p/(1-p), hvor p er sandsynligheden for at svare rigtigt på et item.

[13] Dikotome opgaver har kun to svarmuligheder, fx ja/nej eller rigtig/forkert

[14] Polytome opgaver har flere delopgaver, hvor eleven kan få 0, 1, 2,…, k rigtige

Den statistiske sikkerhed kan primært forbedres ved at øge antallet af opgaver den enkelte elev skal besvare. Antallet af opgaver hænger sammen med tiden til den enkelte test. I den sammenhæng kan det tilføjes, at antallet af point (lig med antal delopgaver) i folkeskolens digitale prøver i matematik uden hjælpemidler, biologi, geografi m.fl. ligger omkring 50.

For yderligere at vurdere om en statistisk usikkerhed på 0,46 ved 23 dikotome opgaver er stor eller lille, kan følgende hentes fra notatet "Om opgavetyper og usikkerhed i de nationale test" (Svend Kreiner, juni 2017[15]):

"*Tallet 0,45 kan derfor bruges som en benchmark værdi, hvis man både vil vurdere, hvor godt den adaptive algoritme har fungeret for en adaptiv test med 20 dikotome opgaver, og hvor godt en ikke-adaptiv test fungerer for elever med forskellige grader af dygtighed. Det kan for eksempel beregnes, at en ikke-adaptiv test med 20 opgaver, hvor sværhedsgraden er ligeligt fordelt fra -2,5 til +2,5, i bedste fald vil resultere i SEM = 0,54 og i værste fald (for meget dygtige og meget svage elever) med SEM = 0,82. Altså dårligere end en fungerende adaptiv test.*"
"*Eller med andre ord: Hvis den adaptive algoritme fungerer efter hensigten vil usikkerheden på elevdygtigheden i en adaptiv test altid være mindre end usikkerheden i almindelige ikke-adaptive test. Hvor meget mindre afhænger af opgavernes sværhedsgrader og af elevernes dygtighed.*"[16]

Antallet af opgaver og dermed den tid, der afsættes til en test, er helt centralt når den statistiske usikkerhed skal vurderes. Adaptive test giver mulighed for, at mindske denne usikkerhed mest muligt.

Målingerne af elevdygtigheden og den statistiske usikkerhed foregår på en såkaldt logit-skala[17]. På denne logit-skala er den statistiske usikkerhed på elevdygtighederne størst for de dygtigste elever og mindst for elever med en dygtighed på midten af skalaen.

Formidlingen af resultaterne til lærerne har siden starten i 2010 foregået på percentilskalaen, 1-100[18]. Omregning fra den grundlæggende logit-skala til

[15] https://www.uvm.dk/folkeskolen/elevplaner-nationale-test--trivselsmaaling-og-sprogproever/nationale-test/om-de-nationale-test

[16] Svend Kreiner (juni 2017). https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf17/sep/170913-om-opgavetyper-og-usikkerhed-i-de-nationale-test.pdf

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

[18] I formidlingen til elever og forældre bliver resultaterne på 100-skalaen omsat til en femtrins skala: En del under gennemsnittet (1-10), under gennemsnittet (11-35), gennemsnittet (35-65), over gennemsnittet (66-90), en del over gennemsnittet (91-100)

percentilskalaen har nogle uheldige egenskaber. Mange elever har en beregnet dygtighed midt på logit-skalaen med en relativt lille forskel imellem sig. Ved omregning til percentilskalaen vil en given forskel i dygtighed strække sig over mange percentiler på midten og over færre i yderområderne af dygtighedsskalaen. En beregnet statistisk usikkerhed på dygtigheden hos elever med en dygtighed på midten af skalaen vil derfor strække sig over flere percentiler end en tilsvarende statistisk usikkerhed hos elever i yderområderne af dygtighedsskalaen. Derfor fremstår den statistiske usikkerhed på elevdygtigheden formidlet på percentilskalaen størst for elever omkring gennemsnittet, hvilket reelt er i modstrid med den faktiske bagvedliggende statistiske usikkerhed. Formidlingen af resultaterne til elever og forældre foregår på en femtrins skala, hvor netop det midterste interval (gennemsnittet) er bredest. Dette opvejer til dels denne uheldige konsekvens af en omregning til en percentilskala.

Reliabiliteten

Reliabiliteten er et udtryk for testens evne til at rangordne eleverne efter elevdygtighed på korrekt måde.

Reliabiliteten er belyst på forskellig vis af BUVM[19]. I 2016 blev beregnet en såkaldt test-retest korrelation. Beregningerne var baseret på elevers testresultater fra de frivillige test. I den frivillige testperiode er det muligt, at tage den samme test to gange med få dages mellemrum. Gentagelsen af en test skal ske uden, at eleven kan huske det første testforløb og uden, at eleven har lært af den første test eller lært nyt mellem de to testafviklinger. Dette er naturligvis vanskelligt i pædagogiske test herunder i de nationale test. Elevers testadfærd, motivation, koncentration mv. kan desuden spille ind på elevens testresultat. Derfor skal disse test-retest resultater vurderes med stor forsigtighed.

I 2016 foretog BUVM ligeledes test-retest simuleringer, hvor 5.000 elever med forskellig dygtighed fik simuleret et elevforløb i testsystemet to gange. Disse simuleringer er uafhængig af testadfærd og korrelationen mellem de simulerede testresultater er derfor en beregning af den teoretiske test-retest korrelation, som man ville kunne observere, hvis testen fungerede fuldstændigt som forventet. Simuleringerne måler således om testsystemet og tilhørende opgavebank kan genskabe rangordningen af elevernes testresultater.

[19] https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf17/jan/170110-uddybende-bilagsnotat-om-de-nationale-tests-maaleegenskaber.pdf

I nuværende evaluering er ovenstående to mål for reliabiliteten suppleret med et tilsvarende *Person Separation Index*[20,21]. De tre udtryk for reliabiliteten ses i Tabel 2.

Tabel 2 Reliabiliteten i de nationale test

Test	Profilområde	Test-retest[1)]	Simuleringer[2)]	PSI[3)]
Dansk læsning 8. klasse	**Sprogforståelse**	0,66	0,84	0,82
	Afkodning	0,85	0,87	0,84
	Tekstforståelse	0,72	0,88	0,84
Matematik 6. klasse	**Tal og algebra**	0,63	0,89	0,82
	Geometri	0,65	0,86	0,80
	Statistik og sandsynlighed	0,68	0,89	0,83

1) Korrelation mellem elevdygtigheden fra to frivillige test

2) Korrelation mellem elevdygtigheden bestemt ved simuleringer i testsystemet

3) Person Separation Index

Kilde: Styrelsen for It og Læring

Der findes forskellige anbefalinger for niveauet af reliabilitet. I Streiner[22] anføres, at en optimal reliabilitet ikke bør være under 0,70. En anden ofte anvendt vurdering er en reliabilitet på mindst 0,80. Betragtes elevers gentagelser af samme test (test-retest), er en reliabilitet på 0,80, med en enkelt undtagelse, ikke opnået blandt testene angivet i tabellen. Ses på simuleringerne og på Person Separation Index, da er reliabiliteten mindst 0,80 i alle test i Tabel 2.

I 23 ud af 30 profilområder ligger reliabiliteten, i form af Person Separation Index, over 0,80, mens de resterende syv profilområder har en lavere reliabilitet. Specielt er reliabiliteten lav i fysik/kemi.

Den lavere reliabilitet målt ved test-retest metoden kan skyldes flere forhold. Hvis eleverne fx har mistet motivationen eller har afvigende testadfærd i andet testforsøg (retest), da kan det være svært at reproducere samme elevdygtighed som i første testforsøg. BUVM har gennemført analyser af test-retest på en specifik skole, hvor læreren undrede sig over store udsving i nogle af elevernes resultater i to gentagne frivillige test afholdt med syv dages mellemrum i efteråret 2014. Gennemgangen af elevernes testforløb viste, at en stor del af eleverne i andet forsøg besvarede langt flere opgaver uden at anvende længere tid. I elevernes andet forsøg besvaredes næsten 50 procent

[20] Person Separation Index udtrykker forholdet mellem usikkerheden på elevdygtigheden på den ene side og spredningen mellem elevernes dygtighed på den anden side

[21] Karl Bang Christensen, Svend Kreiner, Mounir Mesbah (edt): Rasch Models in Health. Wiley 2013

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

flere opgaver end elever på landsplan i gennemsnit gør. Denne forskel i testadfærd kan betyde, at det er vanskelligt at sammenholde en elevs to test.

Samtidig kan den lavere reliabilitet målt ved test-retest også skyldes en for stor statistisk usikkerhed på den beregnede elevdygtighed.

Analyserne er uddybet i:

- Notat 3. Den statistiske usikkerhed og testenes reliabilitet
  - Bilag 3.1. Statistisk usikkerhed på elevdygtighederne
  - Bilag 3.2. Reliabilitet

Er opgavernes sværhedsgrad korrekt bestemt
I evalueringen af de nationale test skal følgende spørgsmål besvares:
1. Regner de nationale test rigtigt?
1a. Herunder skal det klarlægges om opgavernes sværhedsgrader stadig er korrekt og
1b. om opgaverne passer til Rasch-modellen.

Opgaverne i de nationale test udarbejdes af faglige opgavekommissioner. Opgaverne afprøves efterfølgende af ca. 700 elever på det klassetrin testen er målrettet til. Afprøvning af opgaver med henblik på anvendelse i de nationale test er foregået siden maj 2008. Der har i alt været afholdt 14 opgaveafprøvninger i perioden maj 2008 til januar 2019. Opgaveafprøvningerne foregår som en lineær test. På baggrund af elevernes besvarelser fra opgaveafprøvningen foretages en statistisk analyse, hvor det undersøges om opgaverne passer ind i den eksisterende opgavebank. De opgaver, der passer til Rasch-modellen, bliver tilføjet opgavebanken sammen med opgavernes beregnede sværhedsgrader. Alle nye opgaver, der tilføjes opgavebanken, passer således til Rasch-modellen.

I forbindelse med afprøvningen af nye opgaver til opgavebanken medtages hver gang et antal af de eksisterende og tidligere godkendte opgaver fra opgavebanken. Dette giver mulighed for, at undersøge om disse opgavers sværhedsgrad er ændret siden tidligere opgaveafprøvninger.

Efter opgaveafprøvningen i januar 2018 blev det konstateret, at 8 procent af de genafprøvede opgaver havde ændret sværhedsgrad. Efterfølgende blev deres sværhedsgrader opdateret i opgavebanken. Analyser fra opgaveafprøvningen i januar 2019 viser, at 16 procent af årets genafprøvede opgaver har ændret sværhedsgrad. Disse opgavers sværhedsgrader bliver tilsvarende opdateret i den nye version af opgavebanken.

Analyser viser, at opgavernes sværhedsgrad kan ændres over tid. Nogle opgaver opfattes lettere og andre opfattes sværere i dag, end da opgaverne oprindeligt blev afprøvet. Efter hver opgaveafprøvning bliver opgavernes sværhedsgrad opdateret i opgavebanken.

Jeppe Bundsgaard og Svend Kreiner[23] har ligeledes undersøgt opgavernes sværhedsgrad i én test, dansk læsning 8. klasse. De anvender data fra elevernes besvarelser fra de obligatoriske test afviklet i foråret 2017. Bundsgaard og Kreiner finder afvigelse mellem deres beregninger og beregningerne foretaget på baggrund af opgaveafprøvningerne.

Efterfølgende har STIL foretaget samme beregninger for skoleårene 2009/2010, 2013/2014 og 2017/2018. Beregningerne viser, at der kan være forskel i den beregnede opgavesværhed, når disse baseres på elevbesvarelserne fra de obligatoriske test i forhold til, når opgavesværhederne bestemmes på baggrund af egentlige opgaveafprøvninger. Beregning af opgavernes sværhedsgrad baseret på resultater fra de obligatoriske test anvender data indsamlet i adaptive forløb, mens beregning af opgavernes sværhedsgrad baseret på opgaveafprøvninger anvender data fra lineære testforløb. Beregningerne viser også, at andelen af opgaver, hvor den beregnede sværhedsgrad afviger, ikke ændres markant over tid.

De to metoder til fastsættelse af opgavernes sværhedsgrad giver ikke enslydende resultater for alle opgaver. Analyser viser, at andelen af opgaver med afvigelser, er størst for de middelsvære og svære opgaver.

Analyserne er uddybet i:

- Notat 4. Opgavebanken og opgavernes sværhedsgrad
  - Bilag 4.1. Opgaveafprøvningsperioder
  - Bilag 4.2. Skærmdumps fra RUMM
  - Bilag 4.3. Opgavebankens sammensætning i forhold til opgavernes sværhedsgrad
  - Bilag 4.4. Sammenhæng mellem elevernes dygtighed og opgavernes sværhedsgrad
  - Bilag 4.5. Undersøgelse af link-opgavernes ændrede sværhedsgrad
  - Bilag 4.6. Forskel i opgavernes sværhedsgrad

Kan elevernes beregnede dygtighed fra tre profilområder samles til én vurdering af dygtigheden

I evalueringen af de nationale test skal det afdækkes:

2. Om sikkerheden i målingerne af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder. Herunder skal det følgende klarlægges:

a. Kan det påvises, at profilområderne måler forskellige aspekter af den samme bagvedliggende færdighed?

[23] Jeppe Bundsgaard og Svend Kreiner: Undersøgelse af de nationale tests måleegenskaber. Revideret 2. udgave 2019.

b. Kan testresultaterne fra profilområderne slås sammen og dermed forbedre sikkerheden i testene.

De nationale test tester elevernes dygtighed i udvalgte områder og fag. I hvert fag testes eleverne inden for tre hovedområder, der kaldes profilområder. Elevernes dygtighed beregnes i hvert profilområde ud fra de besvarelser eleven har givet på en række opgaver.

Analyser af elevbesvarelser fra dansk læsning i 8. klasse og i matematik 6. klasse viser, at det med stor sandsynlighed er muligt at kombinere elevens resultater fra tre profilområder til ét samlet resultat.

En samlet beregnet elevdygtighed i hvert fag vil være baseret på ca. 60 opgaver og derfor som udgangspunkt være mere sikkert bestemt end de beregnede elevdygtigheder i de enkelte profilområder.

Et samlet resultat for hver elev i hver test kunne være et supplement til afrapporteringen af resultaterne i hvert af de tre profilområder.

Yderligere analyser skal gennemføres for at undersøge, om det er muligt at samle resultaterne fra tre profilområder i de nationale test. Ligeledes udestår en faglig indholdsmæssig afklaring af muligheden for samling af testresultater fra flere profilområder til ét samlet mål.

Analyserne er uddybet i:

- Notat 5. Samling af testresultater fra flere profilområder

Udarbejdet materiale
Til besvarelse af evalueringens delopgave 1 er der udarbejdet 5 notater med tilhørende bilag:

- Notat 1. Algoritmen i testsystemet og beregning af elevdygtigheden
  - Bilag 1.1. Anvendte skalaer til præsentation af elevernes beregnede dygtigheder
  - Bilag 1.2. Opgavebanken i dansk læsning 8. klasse - sprogforståelse
- Notat 2. De nationale tests måleegenskaber
  - Bilag 2.1. Sammenhæng mellem testresultater og karakterer
- Notat 3. Den statistiske usikkerhed og testenes reliabilitet
  - Bilag 3.1. Statistisk usikkerhed på elevdygtighederne
  - Bilag 3.2. Reliabilitet
- Notat 4. Opgavebanken og opgavernes sværhedsgrad
  - Bilag 4.1. Opgaveafprøvningsperioder
  - Bilag 4.2. Skærmdumps fra RUMM
  - Bilag 4.3. Opgavebankens sammensætning i forhold til opgavernes sværhedsgrad
  - Bilag 4.4. Sammenhæng mellem elevernes dygtighed og opgavernes sværhedsgrad
  - Bilag 4.5. Undersøgelse af link-opgavernes ændrede sværhedsgrad
  - Bilag 4.6. Forskel i opgavernes sværhedsgrad
- Notat 5. Samling af testresultater fra flere profilområder

Alle notater er opbygget ensartet med en sammenfatning, en indledning og et antal afsnit. Til de enkelte afsnit kan der være henvist til bilag med yderligere tabeller og figurer.

De 5 notater inklusiv bilag er samlet i rapporten *Evaluering af de statistiske aspekter ved de nationale test*.