DET NATIONALE FORSKNINGS- OG ANALYSECENTER FOR VELFÆRD

Review af evalueringen af de statistiske aspekter ved de nationale test

Delrapport 1: Evaluering af de nationale test

Peter Rohde Skov og Lasse Hønge Flarup

Review af evalueringen af de statistiske aspekter ved de nationale test – Delrapport 1: Evaluering af de nationale test

© VIVE og forfatterne, 2020

e-ISBN: 978-87-7119-744-0

Arkivfoto: VIVE

Projekt: 301403

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

VIVEs publikationer kan frit citeres med tydelig kildeangivelse.

Forord

Folketinget vedtog i marts 2006 indførelsen af de nationale test. Den første obligatoriske testrunde blev gennemført i foråret 2010. De nationale test var ét blandt flere elementer i et lovforslag fra december 2005 om fornyelse af folkeskolen for at forbedre det faglige niveau blandt eleverne gennem styrket, løbende evaluering i folkeskolen.

De nationale test tjener to formål: De skal fungere som et pædagogisk redskab til lærerne og har derudover et styringsformål rettet mod såvel institutioner, kommuner og på nationalt niveau (Undervisningsministeriet, 2005; 2006).

De nationale test tester eleverne i syv forskellige fag fra 2. til 8. klassetrin, heraf fire obligatoriske fag og tre frivillige fag. Samlet bliver det til10 obligatoriske og op til 32 frivillige test i løbet af et skoleforløb.

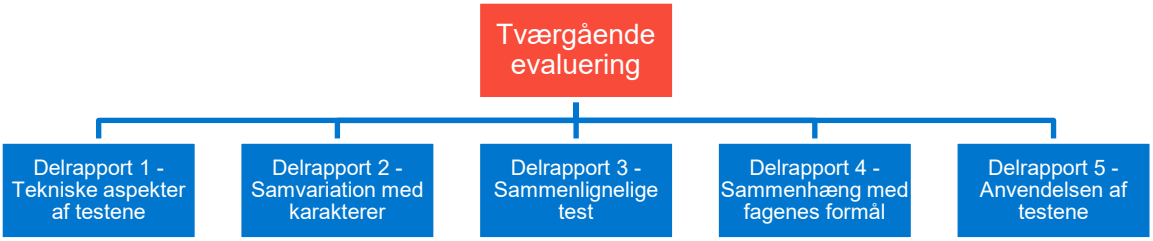
Testene er it-baserede og adaptive, hvilket betyder, at de tilpasser sig den enkelte elev i sværhedsgrad i testforløbet. Hver test består af tre faglige profilområder og er selvscorende. Der gives tilbagemelding per profilområde samt en samlet vurdering. En test kan typisk gennemføres på én lektion, svarende til 45 minutter, med mulighed for at forlænge.

De nationale test blev senest evalueret i 2013. Det blev her besluttet, at der efter en femårig periode igangsættes en ny evaluering. VIVE udarbejder denne nye evaluering af de nationale test.

Evalueringen har til formål at belyse styrker såvel som svagheder omkring indholdet og brugen af de nationale test i folkeskolen samt give et vidensgrundlag, der kan danne afsæt for det fremadrettede arbejde med udvikling og brug af de nationale test i folkeskolen. Evalueringen svarer konkret på det følgende, overordnede evalueringsspørgsmål:

Evalueringsspørgsmål
Har de nationale tests indhold og udformning styrket skolernes evalueringskultur og dermed elevernes faglige niveau?

Evalueringens underordnede undersøgelsesspørgsmål udmønter sig i seks rapporter og en bilagsrapport: én tværgående evalueringsrapport og fem delrapporter, der omhandler hvert sit emne samt en bilagsrapport til delrapport 5.



Hver delrapport besvarer selvstændige undersøgelsesspørgsmål under det overordnede spørgsmål. De fem delrapporter kan, ligesom den tværgående evaluering, læses selvstændigt.

De enkelte delrapporteringer besvarer følgende undersøgelsesspørgsmål:

Del-rapport	Titel	Undersøgelsesspørgsmål
	Evaluering af de nationale test – tværgående rapport	Den tværgående rapport samler resultaterne fra de fem delrapporter
1	Review af evalueringen af de statistiske aspekter ved de nationale test	Har STIL på tilfredsstillende vis besvaret rådgivningsgruppens evalueringsspørgsmål om de nationale tests statistiske usikkerhed, reliabilitet og øvrige måleegenskaber?
2	De nationale tests samvariation med karakterer	Hvad er samvariationen mellem elevers præstationer i testene og karakterer i 8. og 9. klasseprøverne?
3	Kortlægning af sammenlignelige test	Hvilke test findes, der i formål, indhold og omfang minder om de danske nationale test?
4	De nationale tests sammenhæng med fagenes formål	I hvilket omfang er der sammenhæng mellem de nationale test og de centrale dele af faget og fagenes formål jf. Fælles Mål?
5	Anvendelsen af de nationale test samt bilagsrapport med resultater fra spørgeskemaer	Hvordan opleves de nationale test som evalueringsredskab? Hvordan bruges de nationale test i dialogen og opfølgningen på tværs af lokale politikere, forvaltning, skoleledere, lærere, elever og forældre?

Rapporten er Delrapport 1: Review af evalueringen af de statistiske aspekter ved de nationale test. Rapporten gennemgår kritisk STILs evaluering af de nationale test, der vurderer testenes validitet og reliabilitet samt forslag til at forbedre disse. Delrapport 1 er relevant i sammenhæng med Delrapport 2, der ser på, hvor stærk sammenhængen er mellem elevernes resultater i de nationale test og i folkeskolens afgangsprøver. Delrapport 4 ser på sammenhængen mellem opgaverne i de nationale test og målene for de konkrete fag, og Delrapport 5 behandler anvendelsen af resultaterne på både individniveau og aggregeret niveau.

Indhold

Sammenfatning

Styrelsen for It og Læring (STIL) har i forbindelse med evalueringen af de nationale test afrapporteret den statistiske sikkerhed og reliabiliteten i de nationale test. STIL har dokumenteret deres arbejde i en rapport bestående af fem notater fordelt på to dele.

Første del er en validering af den tekniske beregning bag de nationale test, det vil sige spørgsmål om, hvorvidt de nationale test regner rigtigt, om opgavernes sværhedsgrader stadig er korrekte og stadig passer til Rasch-modellen, og om det er muligt at forbedre den adaptive algoritme med henblik på at reducere den statistiske usikkerhed. Den anden del undersøger, hvorvidt målesikkerheden af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder. Dette gøres ved at undersøge, om profilområderne måler forskellige aspekter af den samme bagvedliggende færdighed, og dermed om testresultaterne fra profilområderne kan slås sammen og således forbedre sikkerheden i testene.

VIVE har haft til opgave at reviewe evalueringen gennem nedsættelse og facilitering af en uafhængig gruppe af danske, såvel som nordiske forskere, med særlig viden om test af elever. Forskerne vurderer styrker og svagheder ved resultaterne af STILs dokumentation og analyser af de nationale tests usikkerhed, reliabilitet og øvrige måleegenskaber.

Konkret har VIVE via forskergruppen vurderet styrker og svagheder ved STILs arbejde gennem:

- kritisk stillingtagen til STILs valg af metoder
- kritisk stillingtagen til STILs fund
- kritisk stillingtagen til anvendeligheden af de nationale test på elev og klasseniveau.

Reviewernes kommentarer er grundige og mangfoldige, hvilket også relaterer sig til de forskellige emner, der gennemgås i STILs dokumentation. Reviewene falder i flere kategorier. Der er:

- konstaterende kommentarer, der vedrører STILs fund til overvejelse
- mere eller mindre kritiske kommentarer, der udtrykker et behov for tydeligere forklaringer på analyser eller argumenter for valg af design og metoder
- kommentarer, der udtrykker kritiske punkter eller problemer med STILs dokumentation eller anvendeligheden af de nationale test som testsystem.

Delrapportens resultater

Herunder præsenteres delrapportens fund gennem VIVEs syntese af reviewernes kommentarer til STILs dokumentation for de nationale test.

Reviewerne bemærker overordnet, at STIL har gjort et stort arbejde med at dokumentere de statistiske aspekter af de nationale test, såsom den statistiske sikkerhed og reliabilitet. Der er dog en række områder, hvor der er behov for yderligere forklaringer eller argumentation for valgene, truffet i forbindelse med både selve opbygningen af de nationale test og STILs evaluering af de tekniske aspekter.

Både muligheder og begrænsninger i de nationale test i deres nuværende form

Reviewet viser, at der både er muligheder, men bestemt også begrænsninger, i brugen af de nationale test på elevniveau. De nationale tests opgavers sværhedsgrad er, som dokumenteret

af STIL, beregnet ud fra en lineær test, hvor alle elever får de samme spørgsmål. Brug af lineære test er én måde at fastlægge spørgsmåls sværhedsgrader på. En anden måde er via en adaptiv test, som de nationale test gør brug af i de obligatoriske nationale test, hvor det ikke er spørgsmålenes sværhedsgrader, men elevernes niveau, der fastlægges. STIL viser i sin dokumentation af de nationale test, at opgavernes sværhedsgrad ændrer sig, når man går fra de nuværende lineære test til de adaptive test i opgaveafprøvningen. Reviewerne anbefaler, at alle sværhedsgrader for nationale test skal genberegnes ved at bruge den samme testtype: enten lineære test eller adaptive test. Desuden bør der fokuseres på manglen af svære opgaver, og det bør undersøges nærmere, om der er belæg og behov for tre profilområder for hver af de nationale test, da det mindsker testenes præcision og giver usikre resultater til eleverne, forældre og lærere om elevernes faktiske faglige dygtighed.

Overordnet viser reviewet af STILs notater ligeledes, at der behov for yderligere forklaringer og argumentation for en del af de valg, som er truffet i forbindelse med konstruktionen af de nationale test. Det vil sige, at der efterspørges grundigere beskrivelser, så det i højere grad vil være muligt for en ekstern reviewer at vurdere notaternes korrekthed i forhold til fund og metoder. Det betyder også, at det i en række tilfælde ikke har været muligt for forskergruppen at vurdere indholdet af dokumentationen tilstrækkeligt. De konkrete behov er beskrevet løbende i gennemgangen af reviewet.

Opgaverne vælges på den rigtige måde, og elevdygtighederne beregnes korrekt

I *Notat 1* dokumenterer STIL, at opgaverne vælges på den rigtige måde, og at elevdygtighederne beregnes korrekt. Denne konklusion betyder, at der ikke er tale om en programmeringsfejl i beregningerne. Notatet påviser udelukkende dette ene faktum, og altså ikke om sværhedsgraderne er korrekte, om usikkerheden på målingerne er tilstrækkeligt små, eller om elevernes resultater bliver korrekte.

Reviewerne påpeger, at der er behov for yderligere forklaringer på en række områder i notatet, herunder argumentation for STILs brug af statistiske metoder i evalueringen samt de nationale tests anvendte skalaer til præsentation af elevernes resultater. Dertil er der usikkerhed om argumentet for den valgte adaptive algoritmes valg af opgaver.

Sammenhæng mellem resultater i de nationale test og afgangsprøverne

I *Notat 2* dokumenterer STIL, hvorvidt elevernes resultater fra de nationale test stemmer overens med elevernes resultater fra andre tilsvarende test og prøver. Notatet tester dermed de nationale tests kriterievaliditet. STIL viser, at der er sammenhænge imellem resultater på de nationale test i dansk (læsning) i 8. klassetrin og karakterer ved afgangsprøver på 9. klassetrin og resultater af de nationale test i matematik i 6. klassetrin med afgangsprøver i matematik på 9. klassetrin. Ved at opdele elevernes resultater i percentiler på den normbaserede skala og den kriteriebaserede skala finder STIL positive sammenhænge imellem resultater på de nationale test i dansk (læsning) i 8. klassetrin og karakterer ved afgangsprøver i 9. klassetrin og resultater af de nationale test i matematik i 6. klassetrin med afgangsprøver i matematik i 9. klassetrin. STILs resultater indikerer, at de afprøvede test i dansk (læsning) og matematik er kriterievalide.

Reviewerne er enige om, at der er sammenhæng imellem de nationale test og andre relevante test og prøver, og at dette indikerer, at testene er kriterievalide. Reviewerne savner dog en vurdering af styrken af sammenhængen og forklaringsgraden. Reviewerne mener ligeledes, at det bør fremgå tydeligere, at der alene er tale om kriterievaliditeten, så det ikke forveksles med andre typer af validitet.

Reviewerne er kritiske over for, at der alene fokuseres på to test af de nationale test (dansk (læsning) i 8. klasse og matematik i 6. klasse), og at der ikke er foretaget statistiske test af sammenhængene. Reviewerne stiller derfor forslag til yderligere test, der kan underbygge STILs analyser. Læs endvidere VIVEs delrapport 2 for yderligere test af de nationale tests kriterievaliditet for de øvrige obligatoriske test.

Ifølge reviewerne bør man overveje, om den nuværende omregning af resultaterne til en percentil-skala er formålstjenstlig, da det afleder paradoksale resultater, hvor resultaterne er mest sikre i enderne, men usikre i midten. Og reviewerne mener, at man bør overveje den form, hvormed man formidler elevernes resultater til forældrene i forhold til disse skalaer.

De nationale tests usikkerhed på elevniveau

I *Notat 3* dokumenterer STIL den statistiske usikkerhed på de beregnede elevdygtigheder. STIL finder, at den gennemsnitlige statistiske usikkerhed på elevernes estimerede dygtighed er 0,46 logit. Den gennemsnitlige statistiske usikkerhed er mindst i fysik/kemi i 8. klasse og størst i matematik i 8. klasse. Reviewerne skriver, at notatet har en række uklare punkter, som med fordel kan gøres tydeligere for læseren. Det gælder eksempelvis argumentation for relevansen af de valgte sikkerhedsintervaller og forklaringer på, hvordan beregningerne af dem er gennemført.

Reviewerne finder, at der er mangelfuld argumentation for valget af den statistiske usikkerhed på elevdygtigheden (SEM). Den er fastsat til 0,55, men der er ikke argumenteret tilstrækkeligt for denne værdi, ligesom der ikke bliver reflekteret over, hvad SEM bør være, når der er tale om en pædagogisk test, som de nationale test er.

Usikkerheden på elevniveau hænger sammen med antal opgaver, som eleverne løser og bliver stillet, og dermed testtiden. STIL foreslår i Notat 3, at den statistiske sikkerhed kan forbedres ved at øge testtiden, sådan at eleverne får mulighed for at få flere opgaver, eller ved at tilføre flere polytome opgaver til opgavebanken. Polytome opgaver er opgaver, hvor der er flere delspørgsmål, der tilsammen kan udtrykke, om eleven har svaret rigtigt på hele opgaven eller kun dele – i modsætning til dikotome opgaver med eksempelvis ja/nej-svar.

Reviewerne stiller sig kritiske over for brugen af flere polytome opgaver til at forbedre de nationale tests præcision og mener, at man bør undersøge de nuværende nationale test yderligere, inden der tilføjes eller ændres yderligere ved testene.

Antallet af svære opgaver bør øges for at forbedre præcisionen

Et yderligere forslag fra STILs dokumentation af de nationale test er at tilføre flere svære opgaver. Dette er reviewerne enige i, da dokumentationen for nuværende viser, at der er for få svære opgave, og at dette påvirker den statistiske usikkerhed. En sådan forøgelse vil forbedre de nationale tests præcision.

I *Notat 4* dokumenterer STIL, hvor mange opgaver der er i opgavebanken, hvordan opgaver afprøves, og hvordan besvarelserne fra opgaveafprøvningerne statistisk analyseres. STIL viser, at der er mangel på svære opgaver til de dygtigste elever i flere af profilområderne. STIL finder endvidere, at der er forskel på opgavernes estimerede sværhedsgrad, når disse beregnes på baggrund af de adaptive testforløb (obligatoriske test), og når de beregnes i lineære afprøvningsforløb (opgaveafprøvning).

Reviewerne finder, at metoderne til fastsættelse af opgavernes sværhedsgrader bør undersøges nærmere, da der er stor forskel på opgavernes sværhedsgrad, afhængig af om de er fra

lineære eller adaptive test (som de nationale test er baseret på). Dette betyder, at eleverne og lærerne ikke får den rette information om, hvor dygtige eleverne er i de enkelte fag.

Samling af profilområderne vil øge sikkerheden i målingerne

I *Notat 5* vurderer STIL, om elevernes resultater fra tre profilområder kan samles til ét samlet resultat med en større statistisk sikkerhed, end hvad der er ved de nuværende nationale test. STIL finder, at den statistiske usikkerhed på elevernes estimerede samlede dygtighed i gennemsnit er på ca. 0,30 logit, hvor den i hvert af de analyserede profilområder i gennemsnit ligger på 0,47-0,52 logit. STIL finder derved, at testene kan forbedres ved at supplere de nuværende tre profilområder, der findes for hver af de nationale test, med et samlet resultat fra en samlet skala. Reviewerne er enige i denne betragtning, men reviewerne mener, at der mangler en egentlig test til at understøtte, hvorvidt profilområderne kan sammensættes til et samlet mål for elevernes dygtighed inden for det enkelte fag. Altså om profilområderne måler det samme. Der mangler ligeledes et teoretisk argument for samling af profilområderne til én skala.

Datagrundlag

Rapporten baserer sig på STILs evaluering af de statistiske aspekter af de nationale test bestående af fem hovednotater og tilhørende bilag samt et summary på samlet 159 sider. De fem notater er:

- Notat 1. Algoritmen i testsystemet og beregning af elevdygtigheden
- Notat 2. De nationale tests måleegenskaber
- Notat 3. Den statistiske usikkerhed og testenes reliabilitet
- Notat 4. Opgavebanken og opgavernes sværhedsgrad
- Notat 5. Samling af testresultater fra flere profilområder.

Derudover baseres rapporten på reviewernes afrapportering, hvor de har vurderet styrker og svagheder ved STILs arbejde.

1 Indledning

Indførelsen af de nationale test i folkeskolen blev vedtaget af Folketinget i 2006 som led i en ændring i folkeskoleloven. De nationale test har til hensigt at styrke evalueringskulturen i folkeskolen og sikre en ensartet evaluering af elevernes faglige niveau på tværs af landet med det formål at forbedre det faglige niveau blandt eleverne (Børne & Undervisningsministeriet, 2019).

Baggrund for og indholdet af de nationale test beskrives i større deltalje i den tværgående evalueringsrapport af de nationale test (Flarup, 2020).

De nationale tests statistiske sikkerhed og reliabilitet er centralt for, at den viden, testene bidrager med, kan anvendes troværdigt af lærere, skoleledere, kommuner, politikere og forskere.

Styrelsen for It og Læring (STIL) har i forbindelse med evalueringen af de nationale test afrapporteret den statistiske sikkerhed og reliabilitet i de nationale test. STIL har selv stået for alle beregninger og udarbejdelse af dokumentation af deres arbejde. STILs afrapportering offentliggøres samtidig med VIVEs evaluering af de nationale test. VIVE refererer direkte til STILs afrapportering.[1] STILs rapport består af to dele.

Første del er en validering af den tekniske beregning bag de nationale test. Denne del har til formål at besvare følgende spørgsmål:

- Regner de nationale test rigtigt?
- Er opgavernes sværhedsgrader korrekte og passer de til Rasch-modellen?
- Er det muligt at forbedre den adaptive algoritme med henblik på at reducere den statistiske usikkerhed?

Anden del omhandler, hvorvidt sikkerheden i målingerne af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder. Dette gøres ved at undersøge, om profilområderne måler forskellige aspekter af den samme bagvedliggende færdighed, og dermed om testresultaterne fra profilområderne kan slås sammen og således forbedre sikkerheden i testene.

1.1 Formål

Formålet med denne delrapport er, gennem et eksternt forskerreview, at afdække, om STIL på tilfredsstillende vis besvarer de evalueringsspørgsmål, som er udarbejdet af rådgivningsgruppen til de nationale tests anbefalinger om de nationale tests statistiske usikkerhed, reliabilitet og øvrige måleegenskaber. Rapporten er finansieret af Styrelsen for Undervisning og Kvalitet (STUK).

VIVEs opgave består i at sammensætte en uafhængig gruppe af danske, såvel som nordiske forskere, med særlig viden om test af elever, der har til opgave at reviewe om STILs afrapportering er tilfredsstillende. VIVE præsenterer en syntese af forskernes vurdering. Der er således ikke tale om VIVEs egen vurdering af de pågældende punkter, men om den samlede vurdering fra de tilknyttede forskere i reviewet.

[1] VIVE modtog 31. august 2019 STILs samlede afrapportering. Det er denne version, delrapporten er baseret på.

Medlemmerne af gruppen består af repræsentanter for forskellige forskningsmiljøer, der anvender og udvikler test af elevers faglige præstationer. Forskerne validerer resultaterne af STILs dokumentation og analyser af de nationale tests usikkerhed, reliabilitet og øvrige måleegenskaber. VIVE har udvalgt de fire reviewere, og STUK er blevet orienteret herom. Reviewerne er udvalgt efter deres faglige dygtighed i forhold til at kunne vurdere STILs arbejde, og STUK har ikke gjort indsigelser over for de udvalgte reviewere. Denne delrapport sammenfatter reviewernes kommentarer til STILs dokumentation for de nationale test.

STILs afrapportering består af fem hovednotater og tilhørende bilag samt et summary på samlet 159 sider. De fem notater er:

- 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

De første fire notater omhandler første del, dvs. de nationale tests nøjagtighed. Med andre ord om testene regner rigtigt, om de stadig passer på en Rasch-model, om opgavernes sværhedsgrad stadig er korrekt, og endelig, om det er muligt at forbedre den adaptive algoritme med henblik på at reducere den statistiske usikkerhed. Det femte notat er dedikeret til anden del, der omhandler, hvorvidt sikkerheden i målingerne af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder.

1.2 Design og metode

VIVE tog kontakt til fire potentielle reviewere, som hver især beskæftiger sig med statistiske analysemetoder, som de nationale test er kendetegnet ved, såsom Item Response Theory, Rasch-analyser og psykometriske målingsmodeller. Ud over kendskab til de statistiske metoder, er reviewerne udvalgt, så de repræsenterer forskellige nationale og internationale forskningsmiljøer. Følgende reviewere er udvalgt:

- Anders Holm, professor og direktør, Centre for Research in Social Inequality, Department of Sociology, The University of Western Ontario
- Marie Wiberg, professor, Institut for Statistik, Handelshögskolan vid Umeå Universitet.

- Tine Nielsen, lektor, Psykologisk Institut, Københavns Universitet
- Julius Kristjan Björnsson, forsker, Institutt for Lærerutdanning og Skoleforskning/EKVA, Universitetet i Oslo

Reviewerne blev kontaktet i begyndelsen af juni 2019. Ingen reviewere, der er blevet kontaktet, har afslået at deltage i reviewet. Reviewerne er blevet bedt om at:

- Forholde sig kritisk til STILs valg af metoder og fund
- Forholde sig kritisk til anvendeligheden af de nationale test på elev- og klasseniveau
- Aflevere deres kommentarer skriftligt med en gennemgang af styrker og svagheder ved STILs arbejde.

Reviewerne, der kan læse dansk, modtog den danske version af STILs rapport. Reviewere, der ikke kan læse dansk på et sikkert niveau, fik tilsendt en version af STILs rapport oversat til engelsk. Alle reviewerne fik tilsendt filer med personparametre og sværhedsgrader fra STILs analyser, så de havde mulighed for lave kontrolberegninger. Reviewerne er blevet instrueret i, at STILs notater var fortrolige. VIVE har løbende været i kontakt med reviewerne, hvis der eksempelvis var spørgsmål til reviewet eller til STIL. Hvor det har vist sig nødvendigt, er spørgsmål blevet videresendt til STIL og/eller STUK. STIL og STUK har alene på baggrund af konkrete spørgsmål kommenteret deres rapporter via VIVE og har ikke kommunikeret yderligere med reviewerne.

Reviewerne er ikke blevet informeret om identiteten af de øvrige reviewere, og har, så vidt VIVE er informeret, arbejdet uafhængigt af hinanden.

VIVE har udarbejdet nærværende rapport som en syntese af de fire revieweres kommentarer til de forskellige notater. Rapporten opsummerer og sammenfatter deres kommentarer og bidrager ikke med en selvstændig tolkning af STILs dokumentation.

1.3 Om denne delrapport

Denne delrapport sammenfatter og opsummerer de eksterne revieweres kommentarer til STILs dokumentation. Gennemgangen af notaterne er i den rækkefølge, som STIL har afrapporteret deres resultater i. For hvert notat følger to afsnit. Første afsnit beskriver kort notatets fund og metode. Andet afsnit er syntesen af de eksterne revieweres kommentarer til de enkelte notater samt udklip fra de enkelte reviews.

Syntesen inddeles i positive, neutrale og kritiske kommentarer. Fravær af positive kommentarer er ikke nødvendigvis udtryk for, at STILs dokumentation er mangelfuld, da det også kan være udtryk for, at revieweren ikke har fundet det relevant at kommentere på de elementer af dokumentationen, som er udført tilstrækkeligt og korrekt. Reviewets karakter medfører naturligt et fokus på neutrale kommentarer, eksempelvis ønsker om yderligere uddybning og argumentation og kritiske kommentarer, eksempelvis mangelfulde analyser. Reviewernes kritik kan både omhandle de nationale test, som den kan omhandle STILs valg af metoder til at dokumentere de nationale test.

Der inddrages relevante, udvalgte dele af reviewernes kommentarer. Reviewernes kom-mentarer om eksempelvis fejl i tabelnumre eller kommentarer af formmæssig karakter medtages ikke, medmindre det er af en sådan karakter, at det påvirker rapportens substans. Ligeledes er det heller ikke alle notater, der udgør STILs rapport, der har lige stor vægt hos reviewerne.

Ej heller er det derfor alle notater, der fylder lige meget i denne afrapportering af reviewernes kommentarer.

I det omfang, det er muligt, inkluderes reviewernes uredigerede kommentarer, dog justeret for eventuelle tegnsætnings- og tastefejl. Reviewernes kommentarer er anonymiserede. Det vil sige, at eksempelvis ”Reviewer 1” ikke nødvendigvis stemmer overens med den tidligere opstilling af reviewerne. Referencer til tabelnumrene er til STILs samlede rapport, hvor alle fem notater indgik samlet.

2 Evaluering af de statistiske aspekter ved de nationale test

Dette kapitel sammenfatter evalueringen af de statistiske aspekter ved de nationale test i fem afsnit, målrettet hver af STILs fem notater. Hvert afsnit har to underafsnit.

Første underafsnit beskriver indholdet af STILs notater og er VIVEs gengivelse af de konkrete notater og fund. Andet underafsnit beskriver reviewernes kommentarer. Det består af VIVEs syntese af reviewernes kommentarer samt udsnit fra de konkrete reviews.

Syntesen er opdelt i tre kategorier under hvert notat, positive, neutrale og kritiske vurderinger. Syntesen er udtryk for reviewernes vurdering af det forelagte materiale og de nationale test som helhed. De positive vurderinger er eksempelvis udtryk for konkrete anerkendelser af STILs dokumentation. Typisk vil der ikke være mange af denne slags, da reviewets karakter er at finde udfordringer med det foreliggende materiale. Et fravær af positive vurderinger er således ikke ensbetydende med, at der ikke er udarbejdet tilstrækkelig dokumentation. Neutrale vurderinger har typisk karakter af konstateringer eller behov for uddybninger, som ikke i sig selv er kritiske over for dokumentation eller metode. Kritiske vurderinger har karakter af områder, hvor der stilles konkret kritik af STILs dokumentation eller metode eller af de nationale tests egenskaber. Synteserne er samlet for hver af STILs notater, da disse er organiseret tematisk. I parentes efter vurderingerne fremgår, hvilke reviewere der har samme vurdering. Hvis der ikke fremgår en parentes med angivelse af enkelte reviewere, skal vurderingen læses som værende generel på tværs af de fire reviewere.

De valgte udsnit fra de konkrete reviews er udtryk for de centrale dele af reviewernes kommentarer, der skal være med til enten at eksemplificere konkrete vurderinger eller uddybe disse.

Tværgående for alle notater er følgende pointer:

- Det bemærkes af reviewerne, at STIL har gjort et stort arbejde med at dokumentere de statistiske aspekter af de nationale test, såsom den statistiske sikkerhed og reliabilitet.
- Der ønskes overordnet flere uddybende forklaringer i forhold til, hvilke test STIL har anvendt, og hvilke overvejelser der ligger til grund for valget af disse test.
- Reviewerne kritiserer generelt STILs rapport for, at der anvendes et meget teknisk sprog, der kan være svært tilgængeligt for de fleste læsere. STIL opfordres eksempelvis til at anvende de samme termer igennem afrapporteringen, eksempelvis brug af ordet ”elevdygtighed” i stedet for ”theta” eller ”sværhedsgrad” i stedet for det mere tekniske ”location”.

2.1 Algoritmen i testsystemet og beregning af elevdygtigheden (Notat 1)

Notat 1 beskriver, hvordan opgaverne fra opgavebanken vælges til elevernes testforløb, og hvordan elevernes dygtighed beregnes. De elevdygtigheder, der beregnes i testsystemet, sammenholdes med elevdygtigheder beregnet i et kommercielt softwareprogram.

Formålet er således at vurdere, om elevernes dygtighed beregnes rigtigt i testsystemet.

STIL beskriver i notatet:

- Rasch-modellen
- Beregning af elevdygtigheden
- Valg af opgaver i den adaptive algoritme
- Sammenligning af beregnet elevdygtighed
- Statistisk usikkerhed i testsystem med tilsvarende.

STIL finder i deres notat, at den adaptive algoritme i testsystemet fungerer efter hensigten både i forhold til valg af opgaver fra opgavebanken og i forhold til beregning af elevdygtigheden og den statistiske usikkerhed.

Notatet finder endvidere, at de beregnede elevdygtigheder og tilhørende statistiske usikkerheder i testsystemet i dansk (læsning) i 8. klasse og matematik i 6. klasse fra de obligatoriske nationale test i 2018 er sammenlignet med tilsvarende beregnede elevdygtigheder og usikkerheder ved anvendelse af det kommercielle software-program RUMM. Notatet finder ingen statistisk signifikant forskel mellem elevdygtighederne beregnet i testsystemet og beregnet i RUMM. Den gennemsnitlige forskel er på 0,02 logit. Beregningerne viser overensstemmelse inden for ±0,1 logit mellem elevdygtighederne beregnet i testsystemet og i RUMM for over 99,2 % af alle elevforløb og overensstemmelse inden for ±0,2 logit mellem elevdygtighederne beregnet i testsystemet og i RUMM for over 99,7 % af alle elevforløb.

I notatet viser STIL eksempler på elevforløb for profilområdet "Sprogforståelse" i dansk (læsning) i 8. klasse for at demonstrere algoritmens valg af opgaver. Til notatet er et bilag[2], der viser alle opgaver i opgavebanken i dansk (læsning) i 8. klasse for profilområdet sprogforståelse.

Overordnet set, så viser dette notat, hvorvidt den bagvedliggende algoritme for de nationale test regner rigtigt, og at der derfor ikke er tale om eksempelvis programmeringsfejl. Det skyldes ifølge STIL, at den kommercielle statistikpakke RUMM og algoritmen bag de nationale test kommer frem til stort set de samme beregnede elevdygtigheder, kaldet theta. STIL har testet, at der ikke er statisk sikker forskel på de beregnede elevdygtigheder, de to statistikpakker finder frem til. De største afvigelser i mellem de to algoritmer kommer til udtryk i mere ekstreme tilfælde, hvor eleverne udelukkende besvarer opgaver, der enten er for svære eller for lette for dem.

[2] Bilag 1.2. Opgavebanken i dansk (læsning) i 8. klasse – sprogforståelse (STIL, 2019).

2.1.1 Reviewernes vurderinger af Notat 1

Positive vurderinger

- Notatet viser, at opgaverne vælges på den rigtige måde, og at elevdygtighederne og usikkerhederne beregnes korrekt.

Neutrale vurderinger

- Notatet viser udelukkende, at der ikke er tale om en programmeringsfejl i beregningerne, og således ikke, hvorvidt om sværhedsgraderne er korrekte, om usikkerheden på målingerne er tilstrækkeligt lille, eller om elevernes resultater bliver korrekte.
- Der efterspørges refleksioner om brugen af forskellige statistiske metoder, der frembringer de forskellige resultater, og hvor meget af forskellene, der kan tilegnes disse metoder (Reviewer 3).
- Det bør være mere klart for læserne, hvorfor elevdygtighed (theta) kan falde i mellem -7 og 7 (Reviewer 3).

Kritiske vurderinger

- De anvendte skalaer til præsentation af elevernes beregnede dygtigheder bør forklares bedre, jf. forklaringen i STILs Bilag 1.1[3]. Grundlaget for den kriteriebaserede skala og den normbaserede skala er ikke beskrevet tydeligt nok i STILs dokumentation. Det er uklart, om forældrene præsenteres for den kriteriebaserede skala. Og det er uklart, om elever og forældre er tilstrækkeligt oplyst om, at den normbaserede skal er en relativ skala og ikke, hvordan eleven har klaret sig på det givne område. (Reviewer 1).
- I forhold til de nationale test, så er det er usædvanligt at anvende den valgte adaptive algoritme, hvor der vælges tilfældige opgaver fra et sværhedsgradsinterval. Den almindelige metode i Item Respons Theory er at bruge Fischer Information. Det ser ud til, at den valgte metode har konsekvenser for sikkerheden ved målingen og ser ikke ud til at fungere optimalt (Reviewer 4).

Udvalgte uddrag fra de skriftlige review

I det følgende fremgår reviewernes kommentarer inddelt efter de emner, der bliver behandlet i STILs Notat 1. Disse kommentarer findes også i kondenseret form i den ovenstående syntese.

2.1.2 Udvælgelse af opgaver i den adaptive algoritme

Generelt har de fire reviewere forholdt sig til den måde, hvorpå opgaver udvælges i den adaptive algoritme, som dokumenteret af STIL.

Reviewer 1 og 2 finder, at notatet viser, at opgaverne vælges på den rigtige måde, og at elevdygtighederne beregnes korrekt.

[3] Bilag 1.1. Anvendte skalaer til præsentation af elevernes beregnede dygtigheder (STIL, 2019).

Reviewer 1 Notat 1 viser, ud fra min vurdering, at der ikke er fejl i selve testsystemet bag DNT. Opgaverne vælges på den rigtige måde af det adaptive system, og dygtigheden beregnes korrekt, hvis der ikke er fejl i item-bankens sværhedsgrader og i registreringen af, om opgaver besvares forkert eller korrekt.

I Notat 1 nævnes desuden, at der er forskel på de maksimum likelihood estimater af elevdygtigheden, som DNT benytter, og de vægtede estimater af elevdygtigheden, som RUMM2030 benytter. Det angives, at der er forskelle mellem de to, og at disse forskelle er størst, hvis eleverne skal besvare opgaver, der enten er for lette eller vanskelige for dem (de ekstreme besvarelser). Det må således betyde, at hvis det adaptive system fungerer efter hensigten, og vælger de rette opgaver, er forskellene beskedne.

Det vises altså ikke i dette notat, om sværhedsgraderne er korrekte, om usikkerheden på målingerne er tilstrækkeligt små, eller om elevernes resultater bliver korrekte, men udelukkende at der ikke er programmeringsfejl. Desuden vises det, at der er beskedne forskelle i testsystemets beregning af elevdygtigheden og softwarepakken RUMMs beregning af elevdygtigheden, når eleverne stilles passende opgaver.

Reviewer 2 Sammenligningen mellem de estimerede elevdygtigheder via de nationale test giver stort set samme resultat som, hvis alternative beregningsalgoritmer anvendes. Dette giver beregningerne af elevdygtighederne stor troværdighed.

Reviewer 4 er lidt mere forbeholden, end de andre reviewere, når det kommer til udvælgelse af opgaver i den adaptive algoritme. Reviewer 4 foreslår desuden, at der anvendes andre metoder end de nuværende anvendte:

Reviewer 4 Metoden til at vælge den næste opgave er lidt usædvanlig, men ser ud til at fungere, selvom det er lidt usædvanligt at bruge tilfældigt valgte opgaver fra et interval som den næste opgave. Den sædvanlige ting i IRT er at bruge såkaldte Fischer Information til at vælge den næste opgave, hvilket ville forbedre testen, men dette kan naturligvis gøres på andre måder som her i DNT-prøverne, hvor det ser ud til at have konsekvenser for sikkerheden ved måling. Tilfældigheden ved at vælge opgaver inden for det valgte interval ser ikke ud til at fungere optimalt, men dette er naturligvis også en funktion af opgavebankens sammensætning, dvs. om der er egnede opgaver til det næste trin i testningen. Organiseringen af selve testen ser ud til at være god med en indkøringsperiode (run-in) før den første færdighedsvurdering. Det er dog et åbent spørgsmål, om det er en god ting at skifte mellem profilområder på den beskrevne måde. Det kan tænkes, at dette kan være forstyrrende, og at det ville være bedre at udfylde hvert profilområde individuelt. Men dette er måske et empirisk spørgsmål, der kan testes. At estimere individuelle færdigheder er også usædvanligt, men ser ud til at fungere ok sammenlignet med RUMM bruger Weighted Maximum Likelihood, som kan kritiseres, men til dette formål er denne metode ganske fin. Det ville have været bedre at bruge en EAP (Expected a Posteriori), men dette er også et spørgsmål om smag. EAP ville sandsynligvis give en lidt smallere fordeling af færdigheder i logit-skalaen og dermed måske mindre forskelle mellem elever.

Reviewer 3 efterlyser en diskussion af metoderne, der anvendes i STILs dokumentation:

Reviewer 3 Der er en refleksion i rapporten om, at forskelle kan skyldes valg af metode, hvilket er korrekt. Men hvorfor diskuteres fordele og ulemper ved de to forskellige metoder ikke?

2.2 De nationale tests måleegenskaber (Notat 2)

Notat 2 omhandler, hvorvidt elevernes resultater fra de nationale test stemmer overens med elevernes resultater fra andre tilsvarende test og prøver. For at teste dette undersøger STIL sammenhængen mellem elevernes testresultat i de nationale test og deres efterfølgende præstation i de relevante dele af standpunktsprøverne i 8. klasse samt i folkeskolens afgangsprøver i 9. klasse. Dette gøres, for at elevernes karakter i dansk (læsning) i skoleåret 2017/2018 sammenholdes med elevernes testresultater i de obligatoriske nationale test i dansk (læsning) i 8. klasse i 2016/2017. Tilsvarende sammenholdes elevernes karakter i matematik uden hjælpemidler i folkeskolens prøve i 9. klasse i 2017/2018 med elevernes testresultater i de obligatoriske nationale test i matematik i 6. klasse i 2014/2015.

STIL finder, at der er en positiv sammenhæng mellem elevernes resultater fra de nationale test i dansk (læsning) og matematik og elevernes karakterer i såvel standpunktsprøverne i 8. klasse som i folkeskolens prøver i 9. klasse.[4] STIL finder endvidere, at de nationale test og folkeskolens prøver når til relativt enslydende vurderinger af elevernes faglige niveau i de områder, hvor der testes. Endelig ser STIL på en tidligere undersøgelse af sammenhængen i mellem elevernes resultater fra de nationale test og resultaterne fra PISA-undersøgelserne og DAMVAD's afrapportering af dette, og her finder DAMVAD, at der er sammenhæng i mellem de nationale test og PISA-2012.

2.2.1 Reviewernes vurderinger af Notat 2

Positive vurderinger

- Notat 2 viser, at der er sammenhænge i mellem resultater på de nationale test i dansk (læsning) i 8. klassetrin og karakter ved afgangsprøver i 9. klassetrin og resultater af de nationale test i matematik i 6. klassetrin med afgangsprøver i matematik i 9. klassetrin.

Neutrale vurderinger

- Det bør fremgå af Notat 2, at der i notatets vurdering af validiteten udelukkende er tale om kriterievaliditet. Det skyldes, at det er den eneste form for validitet, der fremlægges i notatet (Reviewer 1).
- Der bør gøres opmærksom på, at korrelationer i mellem de nationale test og PISA 2012 er lave (Reviewer 1 og 4).
- Man bør overveje brugen af normbaserede skalaer og konverteringen til en percentilskala (Reviewer 4).
- Man bør overveje beskrivelsen af skalaen, der anvendes til forældrene, hvor mellemkategorien er meget stor og yderkanterne små. Det er usikkert, om forældrene forstår de usikkerheder, der er forbundet med de midterste kategorier (Reviewer 4).

Kritiske vurderinger

- Det kritiseres, at der i STILs dokumentation kun fokuseres på to test af de nationale test (dansk/ læsning i 8. klasse og matematik i 6. klasse).
- I forhold til STILs dokumentation, kritiseres det, at der ikke er foretaget statistiske test af sammenhængene, og der opfordres til yderligere test, der kan underbygge STILs analyser. Kriterievaliditeten er en af flere test, der kan udføres for at undersøge, hvor

[4] VIVEs delrapport 2 i evalueringen af de nationale test analyserer ligeledes samvariationen mellem de nationale test og karakterer.

gode testene er til at teste elevernes færdigheder. Denne test er forbundet med en vis usikkerhed, da man må formode, at eleverne lærer noget i den tid, der er gået i mellem de nationale test og folkeskolens afgangsprøver i 9. klasse. Det forhold kan give yderligere usikkerhed for testenes kriterievaliditet.

- Notat 2 viser, at der er en positiv sammenhæng mellem de nationale test og afgangsprøverne. Men der savnes dokumentation af styrken af sammenhængen mellem test og afgangsprøver. Som minimum burde forklaringsgrad afrapporteres. Desuden kan man overveje at angive 95-%-prædiktionsintervaller omkring den estimerede sammenhæng mellem test og afgangsprøver (Reviewer 2).

Reviewer 3 har ikke har haft kommentarer til Notat 2.

Udvalgte uddrag fra de skriftlige review

I det følgende fremgår reviewernes kommentarer inddelt efter de emner, der bliver behandlet i STILs Notat 2. Disse kommentarer findes også i kondenseret form i den ovenstående syntese.

2.2.2 Kriterievaliditet

I forhold til STILs notat, så finder Reviewer 1, at STIL viser, at de nationale test lever op til kriterievaliditeten, og at de nationale test korrelerer med passende nøglevariable, som er karakterer i dansk og matematik i 8. og 9. klasse.

Reviewer 1 Notat 2 dokumenterer, at dansk- (læsning) og matematiktestene er kriterievalide, ved at vise, at resultaterne korrelerer med forhold, som påvirker eller påvirkes af de færdigheder, som testene skal måle. Det er klart, at kriterievaliditeten af test, som DNT skal afprøves, og testene forkastes, hvis de ikke korrelerer med passende udvalgte nøglevariable. Det skal dog også bemærkes, at der skal mere end kriterievaliditet til, før det kan konkluderes, at en test fungerer, som den skal. Kriterievaliditet sikrer nemlig ikke mod systematiske fejl (bias) og garanterer ikke, at testene er tilstrækkeligt nøjagtige til det, de skal bruges til. Kriterievaliditetsdokumentationen i Notat 2 viser således kun, at der på dette niveau af undersøgelsen af validiteten, kan konkluderes, at der ikke er fundet nogen fejl endnu, i DNT dansk (læsning) og matematik. Da kriterievaliditetsundersøgelserne udelukkende er lavet for DNT dansk (læsning) og matematik, savnes der således desuden dokumentation for kriterievaliditeten af de resterende testområder.

Tilsvarende finder man hos Reviewer 2 også, at kriterievaliditeten ser ud til at være passende, men kommer også med forslag om, at STIL kunne sammenligne med lignende test i andre lande.

Reviewer 2 Et centralt spørgsmål er, om de nationale test er lige så præcise som afgangsprøverne, eller om der er en større grad af vilkårlighed ved de nationale test sammenlignet med afgangsprøverne. Både test og prøver beskriver et latent underliggende fænomen. Begge er underlagt målefejl – både pga. validitet og reliabilitet. Spørgsmålet er bare, om testene er ligeså præcise som prøverne. Eftersom hvert profilområde dækkes af cirka 19 items, må usikkerheden inden for hvert profilområde være større end ved afgangsprøven, hvor der ifølge rapporten er cirka 50 items. Usikkerheden på de nationale test i forhold til usikkerheden ved afgangsprøverne kan illustreres på flere måder.

I rapporten er det valgt ved at vise den ikke-parametriske regressionskurve for sammenhængen mellem de nationale test i henholdsvis matematik i 6. klasse og afgangsprøven i matematik uden hjælpemidler og dansk (læsning) i 8 klasse og afgangsprøven i læsning i 9. klasse. Det fremgår, at der er en stærk sammenhæng mellem resultaterne i de nationale test og afgangsprøverne.

Rapporten viser således, at der en sammenhæng, men man savner, især i lyset af den offentlige diskussion af reliabiliteten af de nationale test (at eleverne har store afvigelser mellem resultaterne i test og afgangsprøver), dokumentation af styrken af sammenhængen mellem test og afgangsprøver. Som minimum burde forklaringsgrad afrapporteres.

Desuden kunne man overveje at angive 95-%-prædiktionsintervaller omkring den estimerede sammenhæng mellem test og afgangsprøver.

Man kunne også overveje at undersøge sammenhængen mellem test og eksamensresultater fra andre data, fx det Amerikanske National Educational Longitudinal Study (NELS) for at se, om sammenhængen mellem de danske test og prøver har samme styrke som samme sammenhæng i sammenlignelige udenlandske data.

2.2.3 Samvariation med andre mål

Reviewer 1 og 3 har desuden kommentarer om, at relationen mellem de nationale test og karakterer i 8. og 9. klasse viser, at der er en del variation, der kan hænge sammen med den tid, der er gået, siden de nationale test blev taget, og eleverne fik karakterer i fagene, og at denne tid kan påvirke sammenhængen i mellem de nationale test og karaktererne i 8. og 9. klasse:

Reviewer 1 Det kan også bemærkes, at der i Notat 2 skrives, at DNT og folkeskolens prøver "når til relativt enslydende vurderinger af elevernes faglige niveau". Her mener jeg, at det kunne diskuteres, hvad der ligger i ordet relativt, og at vurderingen af dette nuanceres. Der er for eksempel tale om, at 5 % af elever med utilstrækkelige testresultater i læsning i 8. klasse får 7 til eksamen i 9. klasse, og at 16 % med fremragende læseresultater i 8. klasse får 2 eller 4 til eksamen i 9. klasse. Med hensyn til matematik får 9 % med utilstrækkelige resultater i 6. klasse får 7 eller mere til eksamen. For disse elever fremstår resultaterne ikke enslydende. I tillæg kunne det nuanceres, at der er der er tale om sammenligninger af resultater i 8. hhv. 9. klasse for læsning og i 6. hhv. 9. klasse for matematik. Det burde således forventes, at der ville være forskel mellem DNT og prøverne i begge tilfælde (skolens formål er jo, at eleverne bliver dygtigere), men også at forskellen ville være størst for matematik, da der er gået længere tid. Denne analyse ville have været interessant at se som en del af undersøgelsen af kriterievaliditeten.

Reviewer 3 Endelig forstår jeg ikke helt – hvis øvelsen består i at vise, hvor prædiktive de nationale test er for elevens færdigheder – hvorfor man ikke viser sammenhængen mellem test i 9. klasse og elevens præstationer i prøverne i 9. klasse. Jo længere tid der går mellem test og prøve, jo større er muligheden for uoverensstemmelse mellem test og prøve alene på grund af ændret kompetenceniveau hos den enkelte elev. Især hvis testene anvendes i pædagogisk sammenhæng, må man forvente "regression to mean", fordi især svage elever (elever med lav score i testen) må forventes at blive udsat for en ekstraordinær pædagogisk indsats for at løfte deres kompetencer.

Reviewer 4 mener, at sammenhængen mellem PISA 2012 og de nationale test, en analyse foretaget af DAMAD og som der refereres til i STILs Notat 2, er relativt beskeden, og at den samlede afrapportering af de nationale test, hvor de tre profilområder lægges sammen til ét, og som STIL dokumenterer i Notat 5, vil øge sammenhængen mellem PISA og de nationale test.

Reviewer 4 Forholdet til PISA i 2012 er 0,62, dvs. relativt beskedent. Det forklarer kun 38 % af variationen. Og dette var kun mellem læseforståelse fra PISA og tekstforståelse fra DNT. De andre profilområder fra DNT korrelerede signifikant lavere (tabel 2.5 fra DAMVAD-rapporten) omkring 0,46 og 0,49. For matematik rapporteres lignende tal, den højeste sammenhæng mellem DNT-matematik og PISA-matematik 0,67. Der har også været en sammenhæng mellem kompetenceniveauer, og der er en klar sammenhæng, men også nogle uoverensstemmelser, som naturligvis kunne forventes. Det konstateres, at dette er en konsistent sammenhæng, men dette kan der sættes spørgsmålstegn ved. Dette er bestemt konsistent, men ikke imponerende højt. Her mangler en forklaring på, hvad dette betyder, hvad der ikke passer eller holder sig til, fordi dette ikke er en høj korrelation. Her kan den generelle rapportering fra alle profilområder kombineret sandsynligvis forbedre resultatet. Og så skal man huske på, at både læsning og matematik i PISA er sammensat af profilområder (underskalaer), og at en sammenligning mellem den samlede færdighed og PISA ville være meget mere passende og sandsynligvis ville give et bedre resultat."

2.2.4 Den normbaserede skala

I forhold til den normbaserede skala, så mener Reviewer 4, at der er forbehold over for normbaserede skalaer og konverteringen til en percentil-skala, som de anvendes ved de nationale test. Reviewer 4 foreslår, at en anden skala anvendes, da der ellers er tolkningsmæssige problemer, når resultater afrapporteres.

Reviewer 4 Lidt usædvanlig brug af udtrykket normbaseret skala. Dette betyder normalt, at præstationen på en test konverteres til en anden skala, som man opnår ved at standardisere en test, fx på en repræsentativ population. Dette er den metode, som prøver såsom IQ-tests (WISC-WPPSI) og andre sådanne anvendelser. Derefter udtrykkes resultatet i tal, der kommer fra "normen", dvs. fra det udvalg, der blev valgt til at standardisere prøven. Her handler det ikke om en sådan operation, men om en enkel konvertering til en "percentil" skala. Dette skal måske kaldes skalering eller lignende. Percentilskalaen er ikke normbaseret, selvom den er en "normaliseret" skala. Råscore konverteres oftest til en sådan skala med kendte egenskaber, og det er det eneste, der gøres her. Det sædvanlige er at konvertere lineært til en skala som fx en T-score, der i gennemsnit er 50 og SD 10, eller i gennemsnit 500 og SD 100, som i TIMSS og PISA. Dette er ikke normbaseret, men en skalering af den originale score (logit), der sigter mod at fjerne minustal og gøre skalaer sammenlignelige, dvs. ydeevne fra forskellige områder sammenlignelige.

Det er uheldigt at bruge denne [percentil, red.]-skala til skalerede resultater, hvis den har forskellige størrelser i den forstand, at der er langt mellem både det laveste og det højeste tal og en kort afstand mellem tallene i midten. Naturligvis skyldes dette, at denne "percentil"-skala først og fremmest beskriver antallet af enkeltpersoner/elever, der har hvert resultat, og der er selvfølgelig mange i midten. En lineær skalering, til for eksempel en T-score eller noget lignende kan være mere passende, hvis den samme kompetenceforøgelse ville ligge bag en ændring fra 10 til 20 og fra 50 til 60, hvis dette var en gennemsnitlig T-score på 50. På skalaen "percentil" er der langt mellem 1 og 10, men en kort afstand mellem 40 og 50. Men dette er bare noget at

tænke på. Ved rapportering til forældrene forsøger man at rette op ved at have mellemkategorien meget stor og yderkanterne små. Det er et spørgsmål om, at folk forstår dette, især på grund af usikkerheden omkring de midterste kategorier bliver meget store på denne skala, og det giver indtryk af, at testene er meget usikre.

2.3 Den statistiske usikkerhed og testenes reliabilitet (Notat 3)

Notat 3 redegør for den statistiske usikkerhed på de beregnede elevdygtigheder.

STIL skriver, at den statistiske usikkerhed beregnes i testsystemet, og at denne statistiske usikkerhed omsættes til sikkerhedsintervaller omkring elevdygtigheden. Usikkerheden formidles til lærerne i testsystemet. Sikkerhedsintervallerne omregnes til de forskellige skalaer elevernes dygtighed formidles på. I notatet samles op på tidligere beregninger og suppleres med nye baseret på de seneste obligatoriske test. I notatet præsenteres forslag til forbedringer af de nationale test med henblik på reduktion af den statistiske usikkerhed. Formålet er således, at beskrive den statistiske usikkerhed og testenes reliabilitet samt komme med forslag til forbedringer af de nationale test.

STIL finder følgende i deres beregninger af den statistiske usikkerhed og testenes reliabilitet:

- Den gennemsnitlige statistiske usikkerhed på elevernes estimerede dygtighed er 0,46 logit.
- Den gennemsnitlige statistiske usikkerhed er mindst i fysik/kemi i 8. klasse (0,36 i profilområde 3) og størst i matematik i 8. klasse (0,54 i profilområde 3).
- Generelt er usikkerheden størst for de dygtigste elever.
- 93 % af alle obligatoriske testforløb i skoleåret 2017/2018 blev afsluttet med en statistisk usikkerhed under 0,55 SEM.

STIL foretager endvidere omregninger af den statistiske usikkerhed til den normbaserede skala (1-100). Her finder de, at:

- Længden på 68- og 95-%-sikkerhedsintervallet er på henholdsvis ± 12 og ± 22 point.
- På den normbaserede skala er sikkerhedsintervallerne størst på midten af skalaen. STIL beregner også reliabiliteten via Person Separation Index. Dette indeks måler forholdet mellem usikkerheden på den enkelte elevs dygtighed og spredningen mellem elevernes dygtighed.

STILs analyser viser:

- Reiliabiliteten ligger i intervallet 0,74-0,91 for dansk (læsning), matematik og engelsk, mens den ligger i intervallet 0,66-0,70 for fysik/kemi.
- Reliabiliteten ligger over 0,80 i 23 ud af 30 profilområder og under 0,80 i de resterende syv profilområder.

STIL foreslår, at den statistiske sikkerhed kan forbedres ved at:

- Forlænge testtiden, således at eleven når at besvare flere opgaver. STIL skriver, at hvis testtiden øges, så antallet af opgaver eleverne når at besvare øges fra de nuværende ca. 20 til 40, da kan den bedst mulige statistiske usikkerhed reduceres fra 0,45 til 0,32.
- Øge antallet af polytome opgaver
- Tilføje flere svære opgaver til opgavebanken
- Justere algoritmen i testsystemet, så opgaver med størst mulig informationsværdi vælges.

2.3.1 Reviewernes vurderinger af Notat 3

Neutrale vurderinger

STIL dokumenterer, at sikkerhedsintervallerne på den normbaserede skala er meget brede, og målingerne dermed ganske usikre (jf. tabel 14 i Notat 3). Det dokumenteres, at risikoen for, at en elev fejlplaceres på den kriteriebaserede skala er stor (jf. tabel 15 og 47) i dansk (læsning) og matematik. Afsnittet og tilhørende tabeller kunne have været klarere på flere punkter (Reviewer 1):

- I STILs dokumentation bør relevansen af at have både 68-%- og 95-%-intervaller forklares.
- I STILs dokumentation bør forklares, hvordan sikkerhedsintervallerne for percentil-skalaen er beregnet. Der kan ikke være tale om symmetriske intervaller, sådan som det er angivet, da værdierne ikke kan komme under 0 eller over 100 for de mest dygtige henholdsvis de mindst dygtige elever.
- I STILs dokumentation bør forklares, hvordan sikkerhedsintervallerne i forhold til den kriteriebaserede skala er beregnet, og om det er i forhold til konkrete SEM-værdier.
- Der savnes i Notat 3 dokumentation for risikoen for fejlplacering på den kriteriebaserede skala inden for de resterende testområder.

Kritiske vurderinger

- Test-retest-korrelationerne, som udføres i STILs Notat 3, er relativt lave og svære at forstå. Betydningen af resultatet bør uddybes, og forklaringen gøres lettere at læse.
- Den statistiske usikkerhed på elevdygtigheden (SEM) er sat til 0,55. Det er kritisabelt, at der ikke er argumenteret for denne grænses værdi i STILs dokumentation. Der er heller ikke argumenteret for, hvor mange opgaver der bør fastsættes som minimumsgrænse. Reviewerne mener ikke, at STIL i dokumentationen forholder sig til, hvad SEM bør være i en pædagogisk test, som de nationale test, men kun, hvad den er fastsat til. SEM kommer i analysen i gennemsnit ned på under det valgte stopkriterie på SEM = 0,55, men aldrig ned på 0,30, som er den værdi, der sædvanligvis omtales som et ønskværdigt niveau for SEM på personniveau i faglitteraturen (Reviewer 1).
- I forhold til de nationale test er brugen af percentil-skalaen problematisk at anvende, da den giver paradoksale resultater, hvor resultaterne er mest sikre i enderne, men usikre i midten (Reviewer 4).
- I forhold til STILs forslag om at bruge af flere polytome opgaver bør man forholde sig til, at det ikke kan udelukkes, at en elev hurtigere kan besvare et tilfredsstillende antal binære opgaver. Og det er af denne grund ikke givet på forhånd, at de polytome items er bedre (Reviewer 4).
- Flere polytome opgaver kan være med til at forbedre de nationale tests præcision, men det kan være sværere at lave polytome opgaver. Udvælgelsen af opgaver bør være med udgangspunkt i opgaveinformation i stedet for at tilpasse sig elevdygtigheden (Reviewer 4).
- I forhold til de nationale test bør man undersøge de nuværende nationale test yderligere, fx i forhold til at generere flere svære opgaver frem for at påbegynde helt nye tiltag (Reviewer 1).

Der var ingen relevante, entydigt positive kommentarer til Notat 3.

Reviewer 3 har ikke haft kommentarer til Notat 3.

Udvalgte uddrag fra de skriftlige review

I det følgende fremgår reviewernes kommentarer inddelt efter de emner, der bliver behandlet i STILs Notat 3. Disse kommentarer findes også i kondenseret form i den ovenstående syntese.

2.3.2 Den statistiske usikkerhed på elevdygtigheden (SEM)

Reviewer 1 er kritisk over for brugen af en SEM på 0,55 og spørger generelt ind til årsagen til den SEM, der anvendes til stopværdi i de nuværende test. Revieweren angiver endvidere, at der er større SEM, og dermed usikkerhed om elevernes dygtighed, jo dygtigere eleverne er, og at der også er forskelle over fag, i denne SEM.

Reviewer 1

Side 45 indledes med et citat fra COWIs oprindelige løsningsbeskrivelse, hvor der refereres til en SEM-værdi på 0,6 logits [alle SEM angives på denne skala i det efterfølgende] om et af to mulige stopkriterier (det andet angives som 20 opgaver). Dernæst fremgår det, at det endelige stopkriterie blev en SEM på under 0,55 (eller besvarelse af 30 opgaver). Notatet citerer på side 49 Svend Kreiner for at udtale sig om, hvilket niveau af SEM man kan forvente af en lineær test med 20 opgaver: I bedste fald 0,54 og i værste fald 0,82, og at dette er dårligere end en *fungerende* [reviewers kursivering] adaptiv test, hvor man kan forvente en SEM på omkring 0,46 ved 23 delopgaver. Citatet siger således ikke noget om, hvad sikkerheden bør være, og heller ikke, hvor mange opgaver der derfor bør sættes som minimumsgrænsen, men blot at en større sikkerhed kan forventes med samme antal opgaver i en velfungerende CAT sammenholdt med en lineært administreret test.

Det fremgår dog ikke klart i notatet, hvad argumentet for den valgte SEM-grænse på 0,55 var (det samme gælder det første bud på 0,6), og notatet forholder sig ikke til, hvad SEM bør være i pædagogiske test. En artikel udgående fra den daværende skolestyrelse udgivet i tidsskriftet "Journal of Applied Testing Technology" i 2011, angiver, at SEM-stopkriteriet var 0,3. Skyldes dette, at der oprindeligt var valgt et SEM-niveau på 0,6, derefter et SEM-niveau på 0,3, som så igen er ændret til 0,55? En opklaring af dette samt argumenter for niveauet ville være anbefalelsesværdigt. Det dokumenteres i notat 3, at SEM i gennemsnit kommer ned på under det valgte stopkriterie på SEM = 0,55, men aldrig ned på 0,30, som er den værdi, der sædvanligvis omtales som et ønskværdigt niveau for SEM på personniveau i faglitteraturen – dog oftest uden eksplicitte begrundelser. Et udpluk af den lettilgængelige faglitteratur:

I Wainer's primer om Computerized Adaptive Testing (CAT), hvor det fremgår, at man netop i CATs kan opnå meget præcise målinger for alle netop på grund af konstruktionen, og det bemærkes flere gange, at SEM = 0,3 anses som meget præcist.

- I Linacre's notat angives, at det typiske stopniveau er, når SEM = 0,2 logits i Rasch-baserede CATS.
- I PROMIS (Patient-Reported Outcomes Measurement Information System), som er et stort CAT-baseret system til blandt andet at give klinikere reliable patientbaserede resultater på en lang række sundhedsrelaterede mål, og altså også et system, der skal levere resultater på individniveau, angives SEM-kriteriet at være 0,3.
- I en ny dansk rapport om udviklingen af en talblindhedstest angives da også, at "En tommelfingerregel, som følges flere steder, er, at eleverne mindst skal

besvare et antal opgaver, således at standard error of measurement (SEM) er mindre end ca. 0,3."

At det netop er SEM = 0,3 hænger formentlig sammen med, at hvis variansen af theta er lig med 1, så vil en SEM på 0,3 give en reliabilitet på 0,9 (jf. Svend Kreiners rapport om opgavetyper og usikkerhed i DNT, som der også refereres til i Notat 4), som netop er det reliabilitets-niveau, som en anden del af testlitteraturen anbefaler til målinger, der skal bruges til individniveau.

På side 50 i Notat 3 fremgår det desuden, at SEM afhænger af elevdygtigheden, og at SEM er højere for de dygtigste elever end for de mindre dygtige elever. I tabel 13 samt tabel 34-37 ses dette forhold tydeligt, og faktisk at der er en stigende usikkerhed jo dygtigere eleverne er i langt de testområder, klassetrin og profilområder. Værst står det til i engelsk, hvor usikkerheden i profilområde 4 stiger med 0,25 logits (fra 0,39 til 0,66) fra de mindst til de mest dygtige elever. Årsagen til dette angives som muligvis værende mangel på svære opgaver i opgavebanken. Dette er formodentligt korrekt, idet Notat 4 dokumenterer, at der er for få opgaver til de dygtigste elever i mange profilområder (jf. Notat 4).

2.3.3 Sikkerhedsintervaller omkring elevdygtigheden

STIL har i dette notat angivet konfidensintervaller, der angiver sandsynligheden for, at et testresultat er placeret på det sted i de nationale test, som er angivet ved det resultat, der bliver formidlet til elever, forældre og lærere.

Reviewer 1 og Reviewer 4 kommenterer, at brugen af de kriteriebaserede og normbaserede skalaer har problemer, da sikkerhedsintervallerne er meget brede. Det betyder, at det kan være svært at tegne et retvisende billede af elevdygtigheden, når den formidles til lærere, forældre og elever.

Reviewer 1

Notat 3 indeholder også et afsnit om sikkerheds- (eller konfidens-) intervaller, der skal vise, i hvilket interval man med en vis sikkerhed kan sige, at et testresultat vil ligge. Det angives, at der både beregnes 68-%- og 95-%-sikkerhedsintervaller, at de både beregnes på den normbaserede (1-100) skala, men opdelt i 5 intervaller og på den kriteriebaserede (1-6 skala). Disse intervaller angiver således med hhv. 68-%- og 95-%-sikkerhed det interval en given elevscore vil ligge indenfor. Intervallerne angives som længder. Et 95-%-sikkerhedsinterval i midten af den normbaserede percentil-skala med en længde på 28 ifølge tabel 39 betyder, at en elev i 2. klasse, der har scoret 64 i sprogforståelse, med 95 % sandsynlighed har en score mellem 50 og 78, hvilket efter min mening er en potentielt meget stor forskel på, hvad der opnås, og hvad der forventes. Hvis disse formidles til forældrene, kan de formentlig ikke tænke andet, end at scoren for denne elev lige så godt kunne være i det midterste trin som i trinnet over (jf. også tabel 15). 68 % sikkerhedsintervallerne er naturligt kortere end 95-%-intervallerne, til gengæld er sikkerheden, for at elevens præstation ligger inden for intervallet, tilsvarende mindre.

Det dokumenteres, at sikkerhedsintervallerne er meget brede og målingerne dermed ganske usikre (Tabel 14). Det dokumenteres, at risikoen, for at en elev fejlplaceres på den kriteriebaserede skala, er stor (tabel 15 og 47) i dansk (læsning) og matematik. Afsnittet og tilhørende tabeller kunne have været klarere på flere punkter:

Der savnes en forklaring på relevansen af både at have 95-%- og 68-%-sikkerhedsintervaller, da dette ikke synes åbenlyst for mig. Det relevante synes at have sikkerhedsintervaller omkring egentlige scores med inddragelse af konkrete SEM-værdier, idet intervallerne jo afhænger af dette og derfor vil variere med dette.

Der savnes en forklaring af, hvordan sikkerhedsintervallerne for percentil-skalaen er beregnet. Ikke mindst fordi der umuligt kan være tale om symmetriske intervaller, sådan som det er angivet, da værdierne ikke kan komme under 0 eller over 100 for de dygtigste henholdsvis de mindst dygtige elever.

Der savnes tillige en forklaring af, hvordan sikkerhedsintervallerne i forhold til den kriteriebaserede skala er beregnet, og om det er i forhold til konkrete SEM-værdier.

Der savnes dokumentation for risikoen for fejlplacering på den kriteriebaserede skala inden for de resterende testområder.

Reviewer 4 Om det paradoksale resultat, at usikkerheden omkring gennemsnittet forekommer størst, men mindre ved enderne, når der bruges en "percentil" skala, som er det modsatte af, hvad der faktisk sker, når du bruger logit-skalaen. Dette understreger vigtigheden af at revurdere brugen af denne "percentil"-skala. Det bør erstattes af en lineær transformation, fx en T-score eller lignende som nævnt før. Så ville dette problem ikke opstå."

2.3.4 Testenes reliabilitet

STIL henviser i deres Notat 3 til analyser foretaget i 2016, hvor der både beregnes test-retest-korrelationer baseret på gentagne test afviklet som frivillige test og test-retest-korrelationer baseret på simulerede testforløb. Reviewer 1 og 2 bemærker, at test-retest-korrelationerne, som udføres i STILs Notat 3 er relativt lave, sammenlignet med hvad der ellers forventes i test-retest analyser. STIL beskriver i Notat 3, at de lave test-retest kan skyldes, at eleverne lærer at besvare testen og derfor klarer sig bedre.

Reviewer 1 skriver i sine kommentarer, at reliabiliteten flere steder er lavere end den burde være:

Reviewer 1 Fortællingen om test-retest-korrelationerne fra 2016 i starten af afsnittet og tallene herfra i tabel 16 bidrager til at dokumentere, at der i praksis med de virkelige data opnås langt lavere korrelationer, end når disse simuleres (under antagelse af fungerede skalaer). Det bemærkes, at disse tal skal vurderes med stor forsigtighed, men det er jo trods alt tallene.

Reliabiliteten beregnes ved korrelationen mellem simulerede gentagne testresultater og ved PSI (person separation index) for elevdygtigheden. Disse bør være tæt på hinanden, hvilket bekræftes af tallene i tabel 16 og 17.

Ifølge faglitteraturen bør reliabiliteten for test, der anvendes til vurderinger på individniveau, være omkring 0,9 (jf. afsnittet om SEM ovenfor, ville dette svare til SEM = 0,3 logits, hvis variansen på elevdygtigheden var 1) – nogle eksempler:

I Nunnally og Bernsteins klassiker Psykometrisk Teori angives, at minimumsniveauet for reliabiliteten af test scores, hvorpå der baseres vigtige beslutninger [reviewers bemærkning; det vil jeg mene, at DNT-resultater for folkeskoleelever modsvarer], bør være mindst 0,9, mens den ønskværdige standard bør være 0,958.

I en nylig primer om måling med pædagogiske test, angives således, at 0,90 er minimumsniveauet for reliabiliteten, når det handler om beslutninger vedrørende individer, og at 0,95 er den ønskværdige standard.

I en amerikansk underviserguideline angives, at reliabiliteten (tænkt som intern consistent; Cronbach's alpha) i professionelt udviklede standardiserede "high-stakes" tests bør være mindst 0,9, fordi der er tale om enkeltstående testning, hvorpå der drages konklusioner om hver enkelt studerendes niveau på det målte.

Og ikke mindst angives 0,9 som minimumsniveauet for reliabiliteten i IRT-baserede test i den guideline til beskrivelse og vurdering af psykologiske og pædagogiske tests, som er udgivet af den europæiske sammenslutning af psykologforeninger, som også den danske psykologforening er med i.

Det dokumenteres således i Notat 3 med tilhørende tabeller, at reliabiliteten i praksis i de fleste test- og profilområder er lavere, end den burde være, hvis del-testene fungerede optimalt. Særligt lav reliabilitet findes for profilområderne i fysisk/kemi.

Reviewer 2 skriver i sine kommentarer, at det er uklart, hvordan simulationerne, der anvendes i analyserne, er fremkommet. Reviewer 2 skriver endvidere, at det bør undersøges nærmere, hvorfor test-retest-korrelationerne er så lave som angivet i STILs Notat 3. Revieweren angiver, at der bør anvendes metoder, hvor særegne elementer ved de enkelte opgaver holdes konstant, sådan at man undersøger, om eleverne faktisk forbedrer sig, og om dette kan forklare de lave test-retest-korrelationer.

Reviewer 2 Omtalen af test-retest-simuleringer er uforståelige. Læseren (herunder undertegnede) har ingen mulighed for at forstå hvad test-retest-simuleringer er i forhold til test-retest af de frivillige test. Det virker derfor også uforståeligt, hvordan der kan være så stor forskel på test-retest ved de frivillige test og ved simuleringerne. Det er en meget vigtig del af rapporten, da det for brugerne af testene (lærer, elever og forældre) er svært at forstå, hvorfor retest ikke er mere korreleret med test. Det bliver derfor afgørende for rapporten at argumentere for, at test-retest er mindre valide end simuleringerne og det er således meget vigtigt, at læseren dels forstår, hvad simuleringerne går ud på, samt at man i højere grad bør basere vurderingen af reliabilitet på simuleringerne i forhold til test-retest. Jeg forstår og accepterer fint argumentet om, at den lave test-retest-korrelation skyldes, at eleverne i retesten klarer sig meget bedre. Dette kunne måske illustreres med en regressionsanalyse af retest som afhængig variabel og test som forklarende variabel – hvis denne udføres som en lineær sandsynlighedsmodel og med item-specifikke fixed effects (item specifikke dummies) skulle man forvente en regressionskoefficient på over én – da dette indikerer, at sandsynligheden for at svare rigtigt anden gang er større end sandsynligheden for at svare rigtigt første gang. På den måde kan man illustrere, at den lave test-retest-sandsynlighed skyldes, at eleverne forbedrer sig fra test til retest. Man kunne overveje at inkludere elevernes dygtighed (theta) som yderligere forklarende variabel for at undersøge, om det er særlige elever, der ligger bag den lave test-retest-reliabilitet. Man kunne forvente, at især elever i mellemspekteret (på theta) især har udbytte (læring) af at gennemføre den første test.

2.3.5 Forslag til forbedringer af den statistiske sikkerhed

STIL slutter notatet af med en række forslag til forbedring af den statistiske sikkerhed. Disse forslag går på at forlænge testtiden, således at eleven når at besvare flere opgaver; øge anvendelsen af polytome opgaver, hvor der er en række dikotome opgaver inden for det samme

spørgsmål, og som gives score efter hver korrekt besvarelse; justere den adaptive algoritme, så den baseres på opgaveinformation i stedet for at tilpasse sig elevernes dygtighed; og endelig at øge antallet af svære opgaver.

Reviewerne stiller sig kritiske over for brugen af flere polytome opgaver, og Reviewer 2 skriver:

Reviewer 2 Ved vurderingen af, om øget præcision opnås ved at anvende polytome opgaver i stedet for binære items, bør det inddrages i analysen af den opnåede sikkerhed på elevdygtigheden, at testen foretages over et fastlagt tidsrum (45 min.). Hvis eleverne hurtigere kan besvare et tilfredsstillende antal binære opgaver i forhold til, hvor mange polytime items de kan besvare på samme tidsrum, er det ikke givet på forhånd, at de polytome item er bedst.

Ligeledes mener Reviewer 4, at flere polytome opgaver kan være med til at forbedre de nationale tests præcision, men påpeger samtidig at det kan være sværere at lave polytome opgaver. Reviewer 4 angiver endvidere, at udvælgelse af opgaver på baggrund af opgaveinformation i stedet for at tilpasse sig elevdygtigheden vil være en god idé, da denne praksis anvendes i andre adaptive test.

Reviewer 1 mener derimod, at man i stedet for at påbegynde nye tiltag bør undersøge de nuværende nationale test yderligere.

Reviewer 1 Givet rapportens samlede dokumentation, og hvad der kan udledes af den, finder jeg det naturligt at starte et andet sted: nemlig med at undersøge dimensionaliteten henover profilområder, og for så vidt, at profilområder ikke udgør forskellige dimensioner, så score disse samlet. Som beskrevet i forbindelse med Notat 5, vil dette kunne bringe SEM ned til et mere passende niveau, uden at eleverne nødvendigvis skal besvare flere opgaver og dermed øge testtiden. Dernæst vil min anbefaling være, at der fokuseres på at generere flere svære opgaver, hvor disse mangler, da det vil øge sikkerheden for de dygtigste elever (jf. Notat 4). Spørgsmålet er, om det derefter er nødvendigt med yderligere tiltag?

2.4 Opgavebanken og opgavernes sværhedsgrad (Notat 4)

Notat 4 redegør for, hvor mange opgaver der er i opgavebanken, hvordan opgaver afprøves, og hvordan besvarelserne fra opgaveafprøvningerne statistisk analyseres.

Opgavebankens sammensætning af opgaver i forhold til opgavernes sværhedsgrad og i forhold til elevernes dygtighed beskrives. STIL undersøger endvidere i dette notat, om opgavernes sværhedsgrad ændres over tid. Endelig fremlægges forskellige metoder til fastlæggelse af opgavernes sværhedsgrad samt betydningen for elevernes beregnede dygtighed.

Opgaveafprøvningen foregår som en lineær test, hvor eleverne får 2-3 sæt på ca. 30 opgaver i hvert sæt. Et sæt af opgaver kan besvares på 45 minutter.

Ud over nye opgaver til opgavebanken medtages endvidere et antal af de eksisterende og tidligere godkendte opgaver fra opgavebanken. STIL skriver, at de medtager disse allerede eksisterende opgaver for at sikre et overlap mellem blokkene af opgaveafprøvninger, således at nye opgavers sværhedsgrad kan indplaceres på den eksisterende skala. Disse overlapningsopgaver kaldes link-opgaver.

Notat 4 finder:

- Alle nye opgaver, der tilføjes opgavebanken, passer til Rasch-modellen.
- Der er mangel på svære opgaver til de dygtigste elever i flere af profil-områderne.

Analyser fra 2018 viser, at under 10 % af de opgaver, der genafprøves i forbindelse med opgaveafprøvningerne, har ændret deres sværhedsgrad over tid. Foreløbige analyser fra 2019 viser, at 16 % af de opgaver, der genafprøves i forbindelse med opgaveafprøvningerne, har ændret deres sværhedsgrad over tid. STIL konkluderer ud fra dette:

- Der er ikke generel tendens til, at opgavernes sværhedsgrad ændres over tid.

Opgaver med statistisk signifikant ændret sværhedsgrad får denne opdateret i opgavebanken.

Notatet finder endvidere:

- Der er forskel på opgavernes estimerede sværhedsgrad, når disse beregnes på baggrund af de adaptive testforløb (obligatoriske test), og når de beregnes i lineære afprøvningsforløb (opgaveafprøvning).
- Forskellen i de beregnede sværhedsgrader er størst for de svære opgaver, hvilket betyder, at forskellen i de tilsvarende beregnede elevdygtigheder er størst for de dygtigste elever.
- Der er forskel i opgavernes estimerede sværhedsgrad, når disse estimeres med metoden anvendt i RUMM, og når der anvendes open-source-statistikpakken R.

2.4.1 Reviewernes vurderinger af Notat 4

Neutrale vurderinger

- I STILs dokumentation er det vanskeligt at se, om det er opgavernes sværhedsgrad, der har ændret sig over tid, eller om der er andre faktorer, såsom det antal opgaver, der

undersøges, eller ændringer i lærernes eller elevernes adfærd, der har betydning for STILs fund.

- Det er usikkert, om den metode STIL anvender til at teste opgavernes sværhedsgrader, er den rette (Reviewer 4).
- I forhold til de nationale test, så er forskellene på, hvorvidt der testes med lineære eller adaptive test, meget centrale for at afgøre de nationale tests præcision.
- STILs brug af software i dokumentationen kan kritiseres, men finder ikke diskussionen om brugen af software specielt relevant. Beregningsmetoden, som de forskellige statistikpakker anvender, bør i stedet diskuteres.

Kritiske vurderinger

- I forhold til de nationale test bør der tilføjes flere svære opgaver til de nationale test. Antallet har konsekvenser for de nationale test.
- STILs metode til at afprøve opgavernes sværhedsgrad er problematisk.
- I forhold til de nationale test, er det kritisk, at de nationale tests sværhedsgrad, målt ved lineære test, afviger fra de sværhedsgrader, der er fremkommet ved de adaptive test, som anvendes i de obligatoriske test (Reviewer 1).
- Det er i STILs dokumentation ikke optimalt at sammenligne en ægte obligatorisk test med en anden test. Det betyder, at der ikke laves en korrekt sammenligning. Det ville være mere interessant at sammenligne items sværhedsgrad i et år af den obligatoriske test (fx 2018) med et andet år i obligatoriske test (fx 2016 eller 2014). Dette ville have vist, om itemets sværhedsgrad ændrede sig over tid eller ej, og man ville have undersøgt to high stake-test, der er givet i samme format med hinanden (Reviewer 3).
- I forhold til STILs dokumentation kan algoritmen, der bruges af RUMM-pakken, være problematisk. Forskellene i elevdygtigheder produceret af henholdsvis RUMM og TAM er slående, hvilket indikerer, at pairwise conditional estimaterne ikke er helt egnede til dette formål (Reviewer 4).

Der var ingen relevante, entydigt positive kommentarer til Notat 4.

Udvalgte uddrag fra de skriftlige review

I det følgende fremgår reviewernes kommentarer inddelt efter de emner, der bliver behandlet i STILs Notat 4. Disse kommentarer findes også i kondenseret form i den ovenstående syntese.

2.4.2 Opgavernes sværhedsgrader

Generelt finder reviewerne, at STIL i deres dokumentation viser, at de nationale test har generel mangel på svære opgaver, og at dette har konsekvenser for testene. Reviewerne mener, at antallet af svære opgaver bør øges for at forbedre de nationale test. Mens Reviewer 3 og 4 er enige i, at der bør tilføres flere svære opgaver, så det er muligt, at skelne mellem dygtige og meget dygtige elever, hvilket også vil medføre større sikkerhed i testene om elevdygtigheden, generelt.

Reviewer 1	I forbindelse med den overordnede gennemgang af opgaveafprøvningen og opgavebanken i Notat 4, vises item maps (person-opgave-plots) for dansk (læsning) i 8. klasse (de 3 profilområder), og det konkluderes, at der mangler svære opgaver i afkodning og tekstforståelse. Det angives desuden, at "Manglen på svære opgaver gør det primært vanskelligt at skelne de dygtigste og de allerdygtigste elever ved

hjælp af testene. Endvidere bliver den statistiske usikkerhed ikke så lille, som den kunne blive, hvis der var tilstrækkeligt med opgaver, der passede til elevernes dygtighed.". Derefter henvises til bilag 4.4, og skrives, at det samme er tilfældet i flere andre testområder. Sammenfattende angives det, "der er mangel på svære opgaver til de dygtigste elever i flere af profilområderne".

Hvis jeg anvender samme skønsmæssige og rent visuelle metode til at undersøge de resterende item maps i bilag 4.4, som rapportens forfattere tilsyneladende har gjort, så når jeg frem til, at der mangler svære opgaver i afkodning og tekstforståelse på alle klassetrin, for matematik (algebra og stat/sand) på 3. og 6. klassetrin, om end i mindre grad, i engelsk (ordforråd og lytning) på 4. og 7. klassetrin, om end noget mindre på 7. klassetrin, mens der tilsyneladende ikke er problemer i fysik/kemi. Dette fremstår for mig som en væsentlig del af testområderne og klassetrinene (over halvdelen af de obligatoriske test), hvor der mangler svære opgaver, og usikkerheden derfor er større end den behøvede at være, og hvor det ikke er muligt at skelne de dygtige elever fra de meget dygtige elever. Dette burde i sig selv give anledning til løftede øjenbryn, og at man derfor:

1. burde undersøge, om dette har været tilfældet i tidligere år
2. fik lavet nye sværere opgaver til alle disse testområder
3. supplerede den rene visuelle inspektion af item maps med en numerisk vurdering af targeting
4. sprogligt rapporterer disse resultater i større detalje, således at det fremgår tydeligt og præcist, hvilke områder og klassetrin det drejer sig om, samt hvor stor en andel af DNT disse udgør, for at give et mere komplet billede af dette, også for læsere, der ikke er eksperter.

2.4.3 Stabiliteten af sværhedsgrader over tid

I forhold til STILs dokumentation af, at opgavernes sværhedsgrad er den samme over tid, så bemærker reviewerne, at det er vanskeligt at se, om det er opgavernes sværhedsgrad, der har ændret sig over tid, eller om der er andre faktorer, såsom det antal opgaver, der undersøges, eller ændringer i lærernes eller elevernes adfærd, der har betydning for STILs fund.

Reviewer 1

I afprøvninger af nye opgaver medtages et antal (typisk 5-10 opgaver, som benævnes linkopgaver) af de eksisterende og godkendte opgaver i opgavebanken, og at netop denne praksis gør det muligt at undersøge, om de genafprøvede opgaver har ændret sværhedsgrad. Tabel 18 viser, at der i 2019 er 10.969 opgaver i opgavebanken. Det angives, at der i 2019-afprøvningen er anvendt 296 link-opgaver til afprøvning af, om sværhedsgraderne har ændret sig (i 2018 blev anvendt 208 linkopgaver). Disse 296 opgaver udgør under 3 % af opgaverne i opgavebanken, og det er således under 3 % af opgaverne, hvor det er undersøgt, om sværhedsgraderne har ændret sig, og foreløbige analyser viser, at det er tilfældet for 16 % af linkopgaverne, mens det var 8 % i 2018. Disse resultater leder til spørgsmål og problemstillinger, som jeg finder relevante at tage op til overvejelse:

Hvorfor undersøges den tidsmæssige stabilitet af sværhedsgraderne for et så lille antal og procentdel af opgaverne? Det synes at være muligt at undersøge for flere linkopgaver, da den resterende opgavebank ikke byttes ud hvert år.

Når det dokumenteres, at sværhedsgraderne ændres i 8 % af (de få) linkopgaver i 2018 og 16% i 2019, er der ingen grund til, at det samme ikke kunne være tilfældet for de resterende opgaver, som går igen over årene.

Når det dokumenteres, at der er en stigning i den procentdel af linkopgaver, hvor sværhedsgraden ændres i 2018 til 2019, så er det nærliggende, at dette kunne skyldes, at der er tale om en akkumulering over tid, idet opgaver, som ikke er undersøgt tidligere, jo har "båret ændringen med sig". Det kunne også være relevant at overveje, om stigningen kunne være en effekt af øget "teaching to the test", og om dette skyldes undersøges nærmere.

Givet ovennævnte, kunne det dokumenteres, hvilken andel af linkopgaver der ændrer sværhedsgrad fra år til år gennem hele perioden fra 2010 til 2019.

Om end resultaterne omkring sværhedsgradernes stabilitet over tid er væsentlige, så synes de sammenlignet med forskellen på lineære og adaptive sværhedsgrader at være et mindre problem, ud fra de tilgængelige oplysninger. Resultaterne omkring forskellen på lineære og adaptive sværhedsgrader (jf. nedenstående afsnit) betyder dog, at hvis sværhedsgraderne skal sammenlignes over tid, så bør sværhedsgraderne for alle år fra 2010 til nu først omregnes.

Reviewerne finder endvidere STILs metode til at afprøve opgavernes sværhedsgrad problematisk. Reviewer 3 skriver, at den anvendte metode, Rasch-modellen, måske er for simpel til analysernes formål, i forhold til at undersøge opgavernes sværhedsgrad. Revieweren efterlyser endvidere mere information om, hvilke metoder der har været brugt til at undersøge og fjerne opgaver fra opgavebanken.

Reviewer 3

Forslaget om at justere algoritmen er problematisk: Hvis man altid tager det "bedste" items – dvs. de items, der bliver brugt mest, og så udvælger et tilfældigt item blandt lignende items. Hvis man vil have opgavebanken til at leve længe og ikke være kendt for fremtidige testdeltagere, vil det være bedre at have flere items med lignende egenskaber.

I forhold til at fjerne items fra opgavebanken, så står der, at items, der ikke passer på Rasch-modellen, fjernes. Findes der analyser af disse items? Er de nemme eller svære? Spørgsmålet rejses, når de skriver, at de mangler items med høj sværhedsgrad for de dygtige elever (s. 64). Hvordan vil man inddrage flere svære items?

Det er endvidere uklart, hvorfor de [STIL, red.] kun anvender Rasch-modellen til at undersøge model fit af items. Hvis et item ikke passer på Rasch-modellen, så fjerner de det. Det betyder, at de nogle gange fjerner mange items. En årsag til et dårligt fit kan være, at item'et skal modelleres med en mere kompleks model til to-parameter (2PL) eller tre-parameter (3PL) logistiske modeller.

En af rapportens styrker er den adaptive test, dvs. hvordan de vælger det næste item til testdeltagere. Hvis man tænker på fremtiden: flere testdeltagere besvarer flere items, hvis de gentager testen. Muligvis skal man lade fremtidige testdeltagere øve sig i adaptive test forud for testen, hvis det ikke allerede sker.

En begrænsning er, at rapporten ikke altid beskriver, hvilken metode der bruges. Der findes en stor mængde DIF-metoder. Jeg kunne dog ikke finde et eneste sted i rapporten, hvor de beskriver, hvilken metode de bruger, og hvorfor de har brugt den DIF-metode.

Det er en begrænsning, at de kun bruger link-opgaver fra midten af sværhedsgradsskalaen (s. 68). Det er problematisk, da det gør det svært at undersøge, om meget lette eller meget svære opgaver har den samme sværhedsgrad over tid.

En anden begrænsning er spredningen af sværhedsgraderne for link-opgaver. De bør overveje at bruge nogle sværere items som link-opgaver, da de har en tendens til at variere mest, når de sammenligner sværhedsgraderne mellem opgaveafprøvning og obligatoriske test. (Bemærk, dette er også tilfældet for lettere items i profilområdet 1. Figur 15 s. 66).

Reviewer 4 er også i tvivl om den metode, som STIL anvender til at teste opgavernes sværhedsgrader, er den rette.

Reviewer 4 Der bruges 5-10 linkopgaver i hver estimation, i alt 208 linkopgaver. Disse er sandsynligvis kalibreret sammen og en DIF beregnet mellem år. Men selve linkingsmetoden er ikke specificeret. De foregående parametre bruges ikke til at kalibrere de nye items (FCIP-Fixed Common Item Parameters), og dermed få alle opgaver på samme skala og heller ikke en samkalibrering. Der mangler en beskrivelse af, hvordan man sikrer, at alle opgaver ender på samme skala, når nye opgaver føjes til opgavebanken. Er sværhedsgraden af link-opgaverne fixeret? En "concurrent"-estimation af alle opgaver sammen ville imidlertid være en meget mere sikker metode til dette.

2.4.4 Forskelle mellem lineære og adaptive sværhedsgrader

For reviewerne fremstår forskellene på, hvorvidt der testes med lineære eller adaptive test som værende meget centrale i at afgøre de nationale tests præcision.

I forhold til de nationale test, så er Reviewer 1 kritisk over for at opgavernes sværhedsgrad, målt ved lineære test, afviger fra de sværhedsgrader, der er fremkommet ved de adaptive test, som anvendes i de obligatoriske test. Ifølge Reviewer 1, så betyder det, at kritikken, der har været rejst af muligheden for at sammenligne de nationale test over tid, muligvis hænger mere sammen med forskellene i de lineære test og de adaptive test, end det hænger sammen ændringer i sværhedsgraderne over tid. Det udelukker dog ikke at der stadig kan være tidslige problemer med testene. Reviewer 1 finder også, at de fund, som STIL afrapporterer kan dække over tidligere ukendt viden om forskelle mellem lineære og adaptive computerbaserede test. Reviewer 1 skriver således:

Reviewer 1 Startende med afsnittet ”forskellige metoder … ” side 65, dokumenteres det med al tydelighed, at der er forskel på sværhedsgraderne, afhængigt af om de er beregnet på grundlag af opgaveafprøvningerne, som er lineært administrerede test eller på grundlag af de adaptive obligatoriske testforløb. Figur 15 viser tilsyneladende forskelle på personparametre estimeret ud fra lineære hhv. adaptive sværhedsgrader. At jeg skriver tilsyneladende, skyldes, at der i figur 15 angives, at det er theta-værdier (altså elevdygtigheder, der vises), men der i teksten skrives om forskelle i selve sværhedsgraderne, men med en forkert figurhenvisning (figur 4). Jeg går ud fra, at der er tale om forskelle i sværhedsgrader, og at det blot er akserne, der har forkert label, og så er tendensen den samme, som findes i Bundsgaards og Kreiners nylige rapport, hvor de undersøger dette for 2017-data. Dette kunne tyde på, at det, som Bundsgaard og Kreiner finder, i virkeligheden handler om forskelle forårsaget af, om sværhedsgraderne der anvendes, stammer fra de lineære opgaveafprøvninger eller fra de adaptive obligatoriske test, end forskelle over tid. Dette betyder dog ikke, at der ikke også kan være problemer med den tidslige stabilitet af sværhedsgraderne (jf. det forrige afsnit).

Der gås videre med en analyse af, om forskellene i sværhedsgrader for de lineære opgaveafprøvninger henholdsvis de adaptive obligatoriske testforløb ændres over tid. Der er udvalgt 3 tidsnedslag samt dansk (læsning) og matematik. Tabel 19 samt de

tilsvarende tabeller i bilag 4.6 viser ganske rigtigt, at fordelingen af forskellene mellem de sværhedsgraderne fra de lineære opgaveafprøvninger og de adaptive obligatoriske test er ret konstant. Det dokumenteres til gengæld også, at der er tale om store forskelle mellem de 2 estimater af sværhedsgraderne (tabel 19). Forskelle på 0,5 logits er ikke små forskelle (jf., at for de 5 linkopgaver med signifikant forskellige sværhedsgrader over tid, der vises i bilag 4.5, er forskellene henholdsvis 0.35, 0.36, 0.39, 0.40 og 0.91 logits), og for flere end halvdelen af sværhedsgraderne er forskellene større end 0,5 logits. Op til 30 % af opgaverne har forskelle på mere end 1 logit, hvilket er meget store forskelle. Også disse resultater dokumenterer altså det samme, som Bundsgaard og Kreiner fandt med 2017-data, nemlig at der kan være tale om meget store forskelle. Faktisk dokumenterer det nærværende notat, at omfanget af problemet er større, idet det er tilstede for alle klassetrin i dansk (læsning) og matematik, og både i 2010, 2014 og 2018.

Jeg vil anbefale, at de undersøgelser, der dokumenteres i Notat 4, gennemføres for alle test og profilområder samt alle år for at afgøre, om der er samme resultater for flere områder og alle test, således at konsekvenserne af forskellene for estimationen af elevdygtighederne kan dokumenteres og vurderes.

Mens det er relativt let at lokalisere litteratur, der viser, at der er forskel mellem sværhedsgraderne papir-blyant test of CATs, så er det ikke lykkedes mig, inden for en afgrænset tidsramme, at lokalisere litteratur, der siger noget om forskelle i sværhedsgrader i mellem lineært administrerede computerbaserede test og adaptivt administrerede computerbaserede test.

Dokumentationen af disse forskelle fremstår således som vigtig langt ud over DNT og evalueringen af DNT. Jeg vil derfor også anbefale, at der publiceres en videnskabelig artikel om forskellene i sværhedsgrader fra de to administrationsformater, således at denne viden kan komme CAT-udviklere mv. til gode.

Reviewer 3 forholder sig til, at det ikke er optimalt at sammenligne den obligatoriske test med en anden test, og der derfor ikke laves en korrekt sammenligning:

Reviewer 3

Det er ikke overraskende, at sværhedsgraderne skifter mellem opgaveafprøvning og obligatoriske prøver. Det sker også med andre test. Da opgaveafprøvning er lineære test og ikke obligatoriske, er de ”low stake test”. Det ville have været mere interessant, hvis de i stedet ville have sammenlignet items sværhedsgrad i et år af den obligatoriske test (fx 2018) med et andet år i obligatoriske test (fx 2016 eller 2014) i stedet. Dette ville have vist, om itemets sværhedsgrad ændrede sig over tid eller ej, og man ville have undersøgt to high stake-test, der er givet i samme format med hinanden. Dette ville styrke analyserne af ændringer på vanskeligt niveau. Opgaveafprøvning kontra obligatorisk test ville kun give mening, hvis der ikke er angivet nogen obligatoriske test tidligere (s. 68-73).

Det er ikke overraskende, at den største absolutte forskel i sværhedsgrader er for de lettere og svære items, da en adaptiv test er rettet mod dårlige og gode testtagere, mens der gives en lineær test til testtagere med alle evner (s. 73). I en lineær test kan en person have det godt med sig selv, da en dygtig person vil svare på de fleste emner korrekt. I en adaptiv test vil alle til sidst føle, at de fejler, da de får sværere og sværere items, indtil de svarer forkert. Således får en mere dygtig person kun svære items hele tiden, som er sværere og sværere, og derfor kan de føle mere pres i den adaptive test. En mindre dygtig person får lettere og lettere items, og det løfter derfor presset, og de kan muligvis svare på lettere ting. Selvom items skal være ens for alle grupper, de er prøvet inden for, er det sjældent helt sandt i praksis, og det kan muligvis ske, når testtagere med alle forskellige evner besvarer items i den lineære test, men kun meget

dygtige testtagere får meget vanskelige items i den adaptive test. På samme måde er det kun dem, der ikke er dygtige, der vil få meget lette items.

2.4.5 Forskelle i estimationsmetode

Reviewerne stiller sig endvidere kritiske over for STILs brug af software, men finder ikke diskussionen om brugen af software specielt relevant. Det er derimod den beregningsmetode, som de forskellige statistikpakker anvender, der bør diskuteres:

Reviewer 1 På side 71 angives det, at der er forskel på de metoder, der er anvendt til estimation i nærværende rapport og i Bundsgaards og Kreiners rapport, og at disse derfor ikke er sammenlignelige. Det fremhæves desuden, at STIL har anvendt en kommerciel softwarepakke, mens Bundsgaard og Kreiner har anvendt open source-software. Dette er helt korrekt, men ikke specielt relevant, da hverken det ene eller det andet i sig selv i højere grad borger for korrekte beregninger. Med det sagt, så er det korrekt, at der anvendes forskellige metoder til estimation af sværhedsgrader i forskellige softwarepakker til Rasch-analyser. For eksempel: I RUMM (anvendt til de nationale test) anvendes der pairwise conditional estimation, i TAM (anvendt af Bundsgaard og Kreiner anvendes der marginal maximum likelihood estimation, hvilket også er tilfældet, hvis man anvender softwarepakken SAS, og mens der i DIGRAM, anvendes conditional maximum likelihood estimation.

Reviewer 3 skriver dog, at der mangler en diskussion af den anvendte estimationsmetode.

Reviewer 3 R indeholder en stor mængde R-pakker. Det er ikke R, der giver forskellige resultater fra RUMM – det er TAM-pakken og de valgte estimeringsmetoder. Jeg mangler en kritisk diskussion om, hvorfor TAM blev valgt i stedet for fx ltm, mirt eller pcIRT, se også fx Robinson, Johnson, Walton og MacDermid (2019) til sammenligning med RUMM og R-pakkerne: ltm, eRm, TAM og lordif. Jeg mangler også en diskussion med en motivation for, hvorfor man bruger en bestemt estimeringsmetode frem for en anden.

Reviewer 4 er også kritisk over for algoritmen, der bruges af RUMM-pakken og skriver:

Reviewer 4 Forskellene i elevdygtigheder produceret af henholdsvis RUMM og TAM er slående, hvilket indikerer, at pairwise conditional estimaterne ikke er helt egnede til dette formål. Men dette er noget, der måske ikke har så stor betydning, hvis man fortsætter med at bruge den samme metode og sørger for, at antallet af svar bag hver estimation er stort nok (700 er et godt antal, så længe man er sikker på, at hele dygtigheden er inkluderet i prøven). Her demonstreres det igen, hvor uheldigt det er at bruge en percentil-skala til rapportering. Det bør stoppes.

2.5 Samling af testresultater fra flere profilområder (Notat 5)

Notat 5 vurderer, om elevernes resultater fra tre profilområder kan samles til ét samlet resultat med en større statistisk sikkerhed, end hvad der er ved de nuværende nationale test.

Dette undersøges ved at tage udgangspunkt i dansk (læsning) i 8. klasse og matematik i 6. klasse. Begge test er baseret på obligatoriske test fra skoleåret 2017/2018. Der er ikke foretaget vurderinger for andre test eller år.

Der ses indledningsvist på sammenhængene (korrelationer) i mellem profilområderne inden for de to test og det bemærkes, at der er statistisk sammenhæng i mellem de to tests profilområder. STIL skriver, at de undersøger, om profilområderne kan samles ved at bruge Raschanalyser. Notatet finder, at:

- Besvarelserne fra de obligatoriske test i 2017/2018 viser, at de tre profilområder i dansk (læsning) i 8. klasse godt kan antages at måle forskellige egenskaber af én og samme færdighed. Tilsvarende gør sig gældende for matematik i 6. klasse.
- Den samlede skala for elevdygtighed kan betragtes som supplement til den beregnede elevdygtighed i hvert af de tre profilområder for dansk (læsning) i 8. klasse og matematik i 6. klasse.
- Den statistiske usikkerhed på elevernes estimerede samlede dygtighed er i gennemsnit på ca. 0,30 logit, hvor den i gennemsnit i hvert af de analyserede profilområder ligger på 0,47-0,52 logit.

2.5.1 Reviewernes vurderinger af Notat 5

Neutrale vurderinger

- STIL viser i dokumentationen, at der overordnet set er der en vis ræson i at samle profilområderne til én samlet skala. Det vil nedbringe usikkerheden af resultaterne ved de nationale test. Det er reviewerne enige i. Reviewerne mener dog ikke på nuværende tidspunkt, at det er belæg for at slå profilområderne sammen til én skala i STILs dokumentation. Det skyldes, at der ikke er udført formelle test af denne sammenlægning, der understøtter, at profilområderne kun måler én dimension af elevernes faglige kompetencer.
- I forhold til STILs dokumentation efterspørges teoretiske begrundelser for samlingen af profilområderne til en skala.
- I forhold til de nationale test stilles der spørgsmålstegn ved, hvorvidt det giver mening at anvende adaptive test, hvis profilområderne samles til én test (Reviewer 4).

Der var ingen relevante, entydigt positive eller kritiske vurderinger af Notat 5

Udvalgte uddrag fra de skriftlige review

I det følgende fremgår reviewernes kommentarer inddelt efter de emner, der bliver behandlet i STILs Notat 5. Disse kommentarer findes også i kondenseret form i den ovenstående syntese.

2.5.2 Samling af testresultater fra flere profilområder

Reviewerne finder analyserne af, hvorvidt det kan lade sig gøre at samle testresultaterne fra profilområder, som værende lovende.

Reviewer 1 mangler dog en konkret analyse af, hvorvidt hver af de nationale tests tre profilområder måler én dimension, som elevdygtigheden inden for ét fag. Det vil, ifølge revieweren, forbedre usikkerheden af resultaterne, og det vil være et væsentligt skridt i forhold til at formidle elevdygtigheden til elever, forældre og lærere.

Reviewer 1

Analyserne er lovende, fordi det ser ud til, at det accepteres at læsning og matematik er "samlede" færdigheder og ikke opdelt i profilområder. Hvis dette er tilfældet, vil det kunne løse de problemer med usikkerheden på dygtighedsestimaterne, som er dokumenteret i de foregående noter, idet SEM vil kunne bringes ned på et passende niveau, uden at eleverne samlet set besvarer flere opgaver, og uden at testtiden forøges synderligt.

Det skal dog bemærkes, at der i analyserne ikke ser ud til at være anvendt egentlige test for unidimensionalitet, så jeg vil stærkt anbefale, at sådanne test tillige gennemføres. Jeg vil desuden anbefale, at analyserne og dimensionalitetstest også gennemføres på de resterende testområder.

Jeg har ikke de fag-faglige forudsætninger for at vurdere, om opdelingen i profilområder er passende eller ej. Jeg finder det dog bemærkelsesværdigt, at der er tale om præcis 3 profilområder inden for hvert fagområde. Den sædvanlige fremgangsmåde i testudvikling og validering er at definere, hvilke overordnede færdigheder der skal testes. Dernæst, baseret på eksisterende viden, at lave en eventuel opdeling i delområder, og så lave opgaverne inden for disse. I valideringsfasen undersøges så blandt andet, om den opdeling i delområder, som udviklerne har fundet fagligt begrundet, også er den opdeling, der skal være – altså en konfirmatorisk undersøgelse af dimensionaliteten.

Hvis det viser sig for alle testområder, at der er tale om unidimensionelle skalaer, der måler de overordnede færdigheder (læsning, matematik, osv.), så vil det være muligt at kalibrere til en samlet skala pr. færdighed/testområde, hvor der indgår aspekter som der stilles opgaver indenfor. Det vil som nævnt bringe usikkerheden ned på et acceptabelt niveau, og således vil elevresultaterne blive langt mere præcise. Dette er af stor betydning for brugerne (elever, forældre og lærere), og det er min vurdering, at dette er langt vigtigere end at betragte det som et supplement til de profilopdelte resultater.

Reviewer 2 mener også, at der er potentiale i at samle profilområderne i en enkelt skala. Revieweren mener dog ikke, at der er tilstrækkeligt med test til, at man kan udtale sig om, hvorvidt det kan lade sig gøre.

Reviewer 2

Det er en fremragende idé kun at inkludere opgaver, der passer på Rasch-skalaen. Dette sikrer, at hver profilområde kun måler én og kun én faglig dimension. På dette område er de nationale test bedre egnet som faglig evaluering end folkeskolens afgangsprøve, hvor man ikke er sikker på, hvor mange faglige dimensioner af hver afgangsprøve måler, herunder om de faktisk dækker samme faglige kompetencer over tid.

Man savner dog dokumentation for, at hvert profilområde rent faktisk passer på Rasch-skalaen. Der burde være inkluderet test for item differential functioning (DIF), så man rent faktisk kan konstatere, om der er items, der ligger på grænsen til at passe på Rasch-modellen.

Fordi hvert profilområde er tilpasset Rasch-modellen må en vurdering af samlingen af profilområder (for at opnå større sikkerhed ved bedømmelsen af den enkelte elev) kun give mening, hvis de samlede profilområder også passer på Rasch-skalaen. Ellers er

det uklart, hvad det er der rapporteres. For at vurdere, om profilområderne kan aggregeres til et samlet billede af eleven, bør det således dokumenteres, at det samlede mål for elevkompetencer også følger Rasch-modellen.

Både Reviewer 3 og Reviewer 4 mener heller ikke, at der er tilstrækkelig dokumentation for at profilområderne kan samles i én skala. De efterspørger, i lighed med de to øvrige reviewere, dokumentation for, at profilområderne kan samles til et samlet mål for hver national test. Reviewer 4 stiller endvidere spørgsmålstegn ved, hvorvidt, det giver mening at anvende adaptive test, når profilområderne samles til én test:

Reviewer 3 Analysen af, hvorvidt kun testene kan samles i én Rasch-model, er problematisk, da der antages unidimensionalitet. Ifølge Tabel 21, så er nogle af korrelationerne lave (0,48 og 0,36). Der mangler en teoretisk diskussion af, hvilket teoretisk konstrukt, der vil fremkomme, når alle profilområder samles i ét. Der mangler også unidimensionale test af items, hvor der bruges datareduktionsteknikker, såsom faktoranalyse. Til dette kan en række metoder anvendes (scree plots, parallel analysis, MAP, osv.)

Reviewer 4 En samlet Rasch-model blev kørt og viser at ved at sammensætte opgaverne fra de tre profilområder for hver elev, vil SEM falde med ca. 0,2 fra 0,47-0,52 ned til ca. 0,3. Dette er en betydelig forbedring, men alligevel er det en smule tankevækkende, at her er der ca. 50 opgaver, og måske er fordelen ved at bruge en adaptiv test lidt væk. Men SEM ser ud til at være god til denne metode.

Afsnittet konkluderer, at de tre profilområder i henholdsvis dansk (læsning) og matematik måler forskellige aspekter af den samme færdighed. Dette er baseret på en yderligere Rasch-analyse, hvor 12 ud af 823 opgaver på dansk (læsning) og 6 ud af 1019 opgaver i matematik ikke passer til modellen. Dette ser ud til at være et noget vanskeligt resultat. Det kan tænkes, at disse meget få opgaver har nogle specielle funktioner, der ikke opfylder kriterierne i en samlet model, men her ville det være naturligt at foretage en faktoranalyse eller principalkomponentanalyse for at undersøge, om en eller flere komponenter ligger bag henholdsvis dansk (læsning) og matematik. En multidimensionel model ville være endnu bedre til dette, hvor høje korrelationer mellem profilområder er tilladt, og en sådan model kunne give betydeligt mere sikre svar på, hvorvidt de tre områder i hver prøve kan lægges sammen eller ej. Det er klart, det er vigtigt at gøre dette, hvis usikkerheden i målingen reduceres markant, når alle opgaver bruges sammen. En sådan mIRT-model, hvor der er mange Rasch-varianter, ville være den bedste at bruge her sammen med en konfirmatorisk faktoranalyse.

Litteratur

Bundsgaard, J. & Kreiner, S. (2019). *Undersøgelse af De Nationale Tests måleegenskaber. 2. udgave*. København: DPU - Danmarks Institut for Pædagogik og Uddannelse, Aarhus Universitet.

EFPA (2013). *EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers, version 4.2.6*. Bruxelles: EFPA – European Federation of Psychologist's Associations.

Hale, C. D. & Astolfi, D. (2014). *Measuring Learning and Performance: A Primer. 3rd edition*. Florida: Saint Leo University.

Flarup, L. H. (2020). *Evalueringen af de nationale test. Tværgående evalueringsrapport*. København: VIVE – Det Nationale forsknings- og Analysecenter for Velfærd.

Lindenskov, L., Kirsted, K., Allerup, P. & Lindhardt, B. (2019). *Talblindhedsprojektet. Rapport om udvikling af talblindhedstest og vejledningsmateriale*. København & Roskilde: DPU - Danmarks Institut for Pædagogik og Uddannelse, Aarhus Universitet & Professionshøjskolen Absalon.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Robinson, M., Johnson, A. M., Walton, D. M., MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*(1), 1-12.

Undervisningsministeriet (2005). *Lov om ændring af lov om folkeskolen L101*. København: Undervisningsministeriet.

Undervisningsministeriet (2006). *Lov om ændring af lov om folkeskolen L170*. København: Undervisningsministeriet.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F. & Mislevy, R. J. (2000). *Computerized Adaptive Testing – A Primer. Second Edition*. New Jersey: Lawrence Earlbaum Associates, Inc.

Wandal, J. (2011). National Tests in Denmark – CAT as a Pedagogic Tool. *Journal of Applied Testing Technology*, *12*(1), 1-21.

Wells, C. S. & Wollack, J. A. (2003). *An Instructor's Guide to Understanding Test Reliability*. Wisconsin: Testing & Evaluation Services, University of Wisconsin.

VIVE

DET NATIONALE FORSKNINGS-
OG ANALYSECENTER FOR VELFÆRD