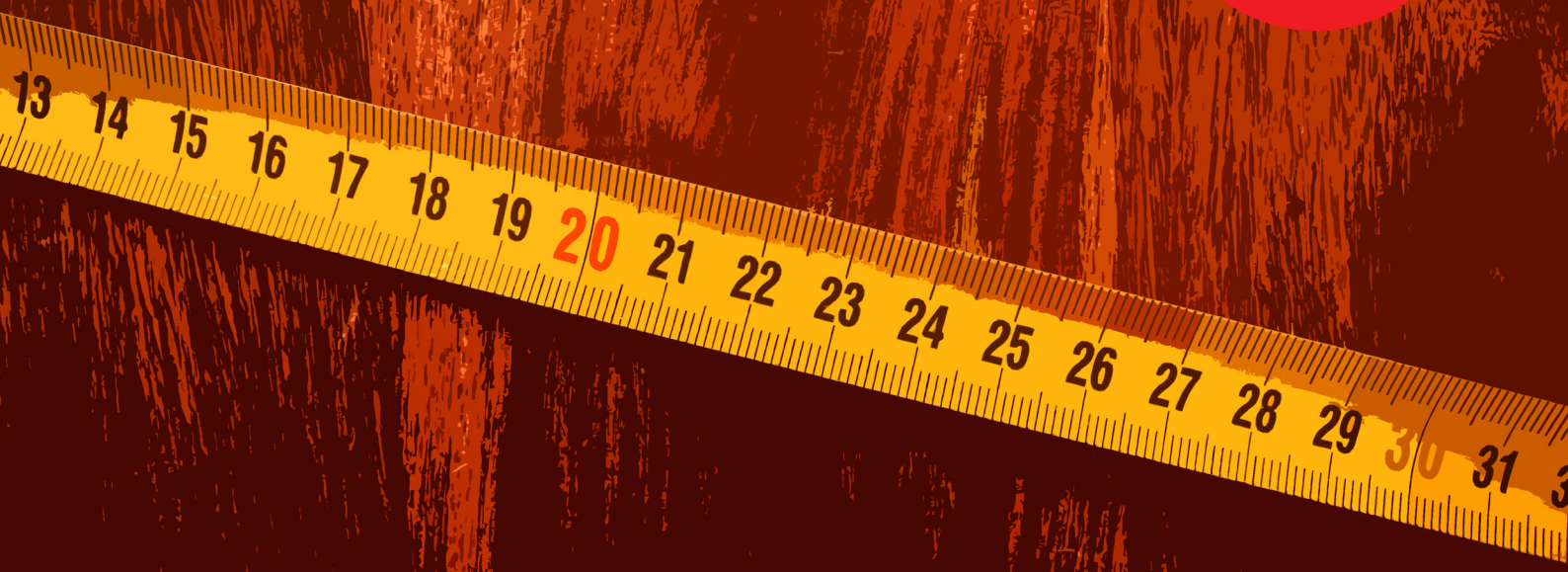
UNDERSØGELSE AF DE NATIONALE TESTS MÅLEEGENSKABER

JEPPE BUNDSGAARD OG SVEND KREINER

REVIDERET 2. UDGAVE



AARHUS UNIVERSITET
DPU

Undersøgelse af De Nationale Tests måleegenskaber

2. udgave

Jeppe Bundsgaard og Svend Kreiner

Aarhus Universitet, Danmarks institut for Pædagogik og Uddannelse
København

Fagfællebedømt af Tine Nielsen og Rolf V Olsen.
ISBN: Elektronisk udgave:
Produceret med R bookdown.
2. udgave Version: 25. april 2019

Indhold

1

Introduktion

1.1 Baggrund

De nationale test har været en væsentlig del af folkeskolens dagligdag siden 2010 og har fået endnu større betydning efter at udviklingen i testresultater blev udpeget som et af succeskriterierne for skolereformen ("Aftale Mellem Regeringen (Socialdemokraterne, Radikale Venstre Og Socialistisk Folkeparti), Venstre Og Dansk Folkeparti Om et Fagligt Løft Af Folkeskolen" 2013). nationale test har været udsat for en del kritik af flere forskellige årsager:

1) Nationale test giver forkerte ikke-valide informationer om elevernes færdigheder (Ravn 2015b; Kousholt 2015).
2) Testresultaterne opfattes som meget usikre blandt andet fordi gentagne målinger af færdighederne vha. nationale test har påfaldende forskelligheder (Riise 2014).
3) Testresultaterne er pædagogisk ubrugelige fordi de er uforståelige og ikke giver oplysninger der er relevante for læreren (Bundsgaard 2018c).
4) Brugen af nationale test som obligatoriske test trækker folkeskolen og fagene i en forkert retning (Bundsgaard 2018b; Bundsgaard og Puck 2016).
5) Afviklingen af testresultaterne presser eleverne unødigt (Bundsgaard og Puck 2016; Wandall, Nørrelund, og Nielsen 2018).
6) Der er mange rapporter om mærkelige og påfaldende testforløb som åbenlyst signalerer at testresultatet må være vildledende og derfor ubrugeligt (Ravn 2015c; Norling 2016).

Punkterne 1, 2 og 6 kan opfattes som spørgsmål vedrørende de måletekniske sider af nationale test. Formålet med det arbejde der beskrives i denne rapport, er at kaste lys over disse problemer ved en analyse af nationale test-resultaterne i dansk, læsning 8. klasse i 2017 og ved en detailanalyse af en række mere eller mindre påfaldende testforløb. Datamaterialet fra 2017 er stillet til rådighed af Undervisningsministeriet mod betaling.

En medvirkende årsag til at vi gik i gang med de mere grundige analyser af nationale tests måleegenskaber og til en række af de konkrete valg af undersøgelsesfoki skyldes matematiklærer på Lemtorpskolen i Lemvig Georg Breddam. Georg Breddam kontaktede Jeppe Bundsgaard i vinteren 2018 for at fortælle om en række observationer han havde gjort i forhold til elevers testforløb. Han havde iagttaget at to tilsyneladende helt ensartede testforløb med samme rækkefølge af rigtige og forkerte svar kunne blive stoppet helt forskellige steder og derved resultere i to helt forskellige resultater. Han havde desuden en mistanke om at svaret på de første tre-fire spørgsmål var afgørende for det endelige resultat. Georg Breddams opsamling af problemstillingen findes i bilag F.

Georg Breddam havde tidligere skrevet til Styrelsen for Undervisning og Kvalitet (STUK) som har ansvaret for nationale test, og delt sin forundring over konkrete forløb som blandt andet kunne rejse tvivl om hvorvidt en god eller dårlig start kunne fastholde eleven på et tidligt estimeret niveau. Svaret fra STUK lød blandt andet således:

> Kort sagt: Dine elevers testforløb ser, i lyset af det adaptive princip, fine ud. De er dog begge eksempler på elever, der enten starter rigtig dårligt eller rigtig godt. Jeg håber, at du kan bruge mine svar. Mange hilsner, NN

Svar af denne art har vi set i flere sammenhænge når lærere har givet udtryk for tvivl om nationale tests funktionsmåde. Da vi ikke mener at det er en fyldestgørende besvarelse på et legitimt spørgsmål, og da Georg Breddams eksempler (som han har opsummeret i bilag F) var overbevisende og tankevækkende, fandt vi det nødvendigt med en grundig og seriøs undersøgelse af om iagttagelserne faktisk var udtryk for grundlæggende problemer, eller om der var tale om enkeltstående afvigelser eller ekstreme tilfælde, og at testen i øvrigt var velfungerende. Georg Breddam formulerer selv sine spørgsmål således (se bilag F):

- Er testen adaptiv i "folkelig forståelse"?
- Når testen beskrives som adaptiv/tilpasser sig den enkelte elevs niveau, er det for mig påfaldende, hvor meget en god/dårlig start sætter sit præg langt hen i testforløbet i de enkelte profilområder - stiafhængighed?
- Hvorledes/hvordan kan testen kategorisere forskellige elever med samme svarmønster vidt forskelligt ved at stoppe forskellige steder i forløbet af items? Nogle elever får mange spørgsmål - mens andre "godkendes" med væsentlig færre spørgsmål i samme profilområde?

Nationale test bygger på en probabilistisk model, og den forudsætter at elever agerer forudsigeligt i forhold til deres dygtighed. Det vil derfor per definition være muligt at finde tilfælde som afviger meget fra det forventede. Denne undersøgelse er derfor blandt andet igangsat for at afgøre om tilsvarende eksempler kan identificeres mere generelt i hele datasættet.

Den anden årsag til at vi har kastet os over de analyser der beskrives i denne rapport, hænger sammen med Undervisningsministeriets igangværende evaluering af de nationale test. Vi deltager begge i en såkaldt ekspertgruppe der rådgiver ministeriet med hensyn til de temaer som en sådan evaluering skal fokusere på. De tekniske sider af denne evaluering er naturligvis et af disse temaer, men vi er bekymrede for at dette tema ikke vil blive taget op på en seriøs måde, fordi der er tale om problemer der kræver en videnskabelig tilgang som ikke kan klares ved hurtige spørgeskemaundersøgelser af den type som de konsulentfirmaer ministeriet normalt benytter sig af, anvender.

Da de måletekniske egenskaber er helt centrale og fordi mange af de problemer som lærere omtaler, kan skyldes sådanne problemer, har det været os magtpåliggende at vise at forskningsbaserede evalueringer ikke nødvendigvis behøver at tage år, men kan klares i løbet af relativt få måneder, selvom der er tale om komplicerede og udfordrende analyser, og selvom det også – som det har været tilfældet i forbindelse med dette arbejde – har været nødvendigt at udvikle metoder til at løse problemer hvor de grydeklare løsninger ikke forelå. Den forskningsbaserede tilgang er årsagen til at det kan lade sig gøre, fordi den sikrer at den nødvendig viden om problemer og metoder er tilgængelige.

Det er af den årsag at vi har foretaget de analyser af den del af nationale test, der beskrives i denne rapport. Resultaterne var ikke givne på forhånd, men vi forventede at rapporten under alle omstændigheder ville være nyttig.

Hvis det viste sig at der ikke var store problemer med testen i læsning i 8. klasse, ville resultaterne være beroligende, selvom nødvendigheden af at se på andre dele stadig ville være til stede. I det tilfælde ville rapporten illustrere hvordan det skal gøres, og at det kan lade sig gøre inden for rimelige tidsrammer hvis opgaven overlades til eksperter der ved hvordan det skal gøres.

Hvis det på den anden side viste sig at der var så store problemer med testen af læsning i 8. klasse at det påvirker testresultater både på elev- og populationsniveau, ville situationen være mere kritisk. Det ville i givet fald være et signal om at læsetesten i 8. klasse ikke kan benyttes i sin nuværende form, og det ville rejse så megen tvivl om resten af de nationale test, at det burde overvejes seriøst at sætte brugen af nationale test i bero indtil de øvrige dele af nationale test var undersøgt.

For ikke at blive misforstået skal vi understrege at rapporten i givet fald kun kan dokumentere at der enten ikke ser ud til at være nævneværdige problemer, eller at nationale test regner forkert i testen af læsning i 8. klasse og leverer forkerte og dermed vildledende testresultater.

Rapporten præsenterer resultaterne af en række statistiske analyser som bygger på forholdsvis avanceret matematik og beregningsteknik. For at give den interesserede læser mulighed for at få en fornemmelse af

grundlaget for analyserne, vil rapporten indeholde kortfattede forklaringer og beskrivelser af psykometri og pædagogiske test i al almindelighed og om nationale test i særdeleshed. Men det skal understreges at målet med denne rapport ikke er at formidle generel viden om hverken psykometri, pædagogiske test eller nationale test. Formålet er kun at dokumentere vores resultater på en måde som eksperter med kendskab til disse emner vil kunne vurdere og tage stilling til. Den interesserede læser vil kunne benytte sig af referencerne sidst i rapporten i det omfang beskrivelserne ikke slår til.

En nærmere diskussion af årsagerne til eventuelle problemer og den måde de i givet fald skal håndteres på, ligger uden for rammerne af dette arbejde, men vi vil – med udgangspunkt i mange års arbejde med udvikling og afprøvning af pædagogiske test – alligevel tillade os at drage vores foreløbige konklusioner og formulere vores anbefalinger i forventning om at det bliver forstået at disse konklusioner og anbefalinger er oplæg til en nødvendig diskussion af hvad der skal ske med nationale test.

Tak til Georg Breddam og de mange andre lærere der har brugt tid og energi på at dokumetere uforklarlige testforløb og uhensigtsmæssigheder ved nationale test.

Tak også til Elisa Nadire Caeli for meget grundig læsning af rapporten og mange gode kommentaerer og forslag samt korrekturlæsning, ikke mindst. Og tak til Christian Christrup Kjeldsen og Morten Rasmus Puck for at give feedback på tidligere udgaver af denne rapport.

1.2 Problemstillingen

Validiteten af testresultaterne forudsætter at testforløbene passer til en psykometrisk skalamodel. Psykometrien omfatter flere sådanne modeller, fx modeller for konfirmatoriske faktoranalyser eller såkaldte IRT-modeller (Item Response Theory). Af disse modeller er den såkaldte Rasch-model den mest interessante model fordi den udover at have en række specielle egenskaber også er den enkleste og lettest forståelige model.

Rasch-modellen der har været anvendt i forbindelse med udviklingen af nationale test, vil blive introduceret i kapitel 3.3. På dette tidspunkt er det tilstrækkeligt at gøre opmærksom på at en såkaldt Rasch-analyse indeholder følgende trin:

1) Estimation af opgavernes sværhedsgrader.
2) Afprøvning af testforløbenes tilpasning til modellen (herunder om der er system i hvilke opgaver eleverne besvarer rigtigt, og hvilke de besvarer forkert).
3) Beregning af elevernes dygtighed ud fra oplysninger om hvor mange opgaver der er besvaret korrekt, og oplysninger om sværhedsgraderne på de opgaver som eleverne har besvaret.

Tilpasning mellem testforløb og model og præcise oplysninger om opgavernes sværhedsgrader er vigtige for alle test, uanset om der er tale om lineære test hvor alle elever besvarer de samme opgaver, eller om adaptive test som nationale test, men kravene er væsentlig mere centrale for adaptive test end for lineære test. Rasch-modellen er som omtalt en probabilistisk model, hvilket betyder at det forventes at elever kan svare rigtigt på opgaver der egentlig er "for svære" for dem, og forkert på opgaver man kunne forvente de kunne svare rigtigt på.

I lineære test kan problemerne være begrænsede fordi alle elever svarer på de samme opgaver, og fordi dygtigheden vurderes ud fra antallet af korrekte svar. Skulle der være en enkelt opgave der falder ved siden af, eller skulle der være problemer med sværhedsgraderne, er det problemer der gælder på samme måde for alle elever. Det vil stadig være de dygtigste elever der har mange rigtige svar, og de mindre dygtige der har færre.

I adaptive test hvor den adaptive algoritme udvælger opgaver ud fra det aktuelle estimat af dygtigheden og ud fra algoritmens viden om opgavernes sværhedsgrader, kan fejl i model og/eller sværhedsgrader være

ødelæggende fordi den adaptive algoritme systematisk vælger forkerte opgaver og beregner forkerte estimater af dygtigheden i løbet af processen.

I forbindelse med adaptive test vil konsekvenserne af disse fejl være mere usikre målinger af dygtighed, og det vil især betyde at der kan være store forskelle på gentagne målinger af den samme elev fordi de fejl der vil være i det første udvalg af opgaver, vil være nogle andre, end dem der er i det andet udvalg.

Selvom den betydelige usikkerhed i nationale test har kunnet bortforklares med at resultater fra pædagogiske test altid forekommer, står det grundlæggende problem tilbage. Giver nationale test mere usikre resultater end man kan forvente, og skyldes det (blandt andet) at der er problemer med den psykometriske skalamodel og/eller de sværhedsgrader som den adaptive algoritme bruger når opgaverne udvælges og dygtigheden beregnes?

De nationale test blev udviklet fra 2007-2010 og sværhedsgraderne og tilpasningen til Rasch-modellen vurderet ud fra et i psykometriske sammenhænge meget stort datamateriale. Dokumentationen af resultaterne (som er den ene af os bekendt) er aldrig offentliggjort.

Siden da er der tilføjet nye opgaver og fjernet gamle opgaver, og sværhedsgraderne er efter sigende (Ravn 2015a) blevet genberegnet i 2014. Vi er ikke bekendt med at der eksisterer dokumentation af denne genberegning.

Der er flere mulige grunde til at opgavernes sværhedsgrader kan have ændret sig siden 2010:

1) Testsituationen i forbindelse med indsamling af testresultater i 2007-2009.
2) Teaching-to-the-test i form af træning af eleverne i at løse den slags opgaver som findes i nationale test.
3) Ændring af elevernes oplevelse af testsituationen fra low-stakes til high-stakes.

Det betyder ikke nødvendigvis at sværhedsgraderne har ændret sig, men det er ikke en selvfølge at de ikke har ændret sig.

Der er derfor grundlag for at analysere aktuelle data fra nationale test og at forsøge at besvare følgende to spørgsmål for at afprøve om nationale test estimerer dygtigheden på den rigtige måde:

1) Lever nationale tests opgaver op til Rasch-modellens krav?
2) Benytter nationale test de korrekte sværhedsgrader?

Og hvis disse spørgsmål ikke besvares positivt at stille spørgsmålene:

3) Hvilke konsekvenser har de identificerede problemer?
4) Er de estimater som nationale test leverer, tilstrækkeligt sikre?

De fire spørgsmål kan siges at være centrale. Hvis de besvares positivt, kan man argumentere for at nationale test generelt fungerer som den skal teknisk set. Hvis det kun er de tre første der besvares positivt, er det nødvendigt at se på årsagerne til at usikkerheden er for stor, og om det evt. skyldes en fejl i programmeringen af den adaptive algoritme, eller at der ikke er tilstrækkelige opgaver til at den adaptive algoritme altid kan finde en opgave der passer til elevens niveau.

Endelig betyder det forhold at nationale test fungerer som den skal, ikke at der ikke kan være problemer i forbindelse med enkelte testforløb. Rasch-modellen er probabilistisk så der er altid en vis sandsynlighed for at tilsyneladende usædvanlige testforløb vil dukke op. Spørgsmålet er derfor ikke om testforløbene er usædvanlige. Spørgsmålet er om de er så usandsynlige at det er vanskeligt at tro på at det kun er elevens dygtighed og sværhedsgraderne af de stillede opgaver der ligger bag testresultatet, således at vurderingen af elevens dygtighed er utroværdig. Den analyse som denne rapport præsenterer, skal også kaste lys over dette spørgsmål og vise hvorledes man kan opdage usandsynlige forløb.

1.3 Plan for analysen

For at besvare disse spørgsmål vil analysen af nationale test-data fra 2017 omfatte følgende:

1) Ny estimation af sværhedsgraderne og en sammenligning med sværhedsgraderne i item-banken
2) En afprøvning af tilpasningen til Rasch-modellen
3) En vurdering af usikkerheden på målingerne ifølge 2017-analysen
4) En analyse af fordelingen af dygtigheden ifølge 2017-analysen og en sammenligning med fordelingen af dygtigheden ifølge nationale tests beregninger
5) En detailanalyse af en række enkeltstående forløb for at illustrere hvorledes den adaptive algoritme fungerer, og for at afprøve om tilsyneladende usædvanlige forløb også er usandsynlige (signifikante), og hvilke konsekvenser det i så fald har for målingerne af dygtigheden.

Spørgsmål 2 er et kompliceret spørgsmål som vil blive besvaret i en efterfølgende rapport. Denne rapport har derfor fokus på de fire øvrige spørgsmål.

1.4 Tilføjelse til 2. udgave

Dette er 2. udgave af rapporten. Denne udgave har vi udarbejdet efter en henvendelse fra Morten Rasmus Puck. Han havde ved læsning af rapporten opdaget at der var en række fejl i den tabel over itemsværhedsgrader der er gengivet i bilag C. Ved gennemgang af vores scripts til produktion af de data der indgår i tabellen, opdagede vi at der var sket en fejl som betød at item-location for polytome items (som kun findes i profilområde 3) var blevet indskrevet i profilområde 1 og derved havde overskrevet de korrekte sværhedsgrader. Det betød at 90 items havde fået forkerte sværhedsgrader i profilområde 1, og at 90 items i profilområde 3 var sat lig med den første kategoris threshold. Profilområde 2 er ikke påvirket af fejlen. Vi har alene anvendt disse værdier i vores illustration af forskellen på itemsværhedsgrader i afsnit 4.3.1, og fejlen vedrører således alene vores behandling af forskelle i itemsværhedsgrader mellem 2017-analysen og DNT-analysen. I bilag G gennemgår vi konsekvenserne af denne fejl. Vi konkluderer på den baggrund at der ikke er grund til at ændre i konklusionerne i rapporten.

Vi vil benytte lejligheden til at takke Morten Rasmus Puck mange gange for at have studerent rapporten så grundigt og identificere den yderst beklagelige fejl.

2

Forfatterne, læsevejledning og resultater

2.1 Forfatterne

Rapportens forfattere har begge været optaget af nationale test fra forskellige perspektiver i en lang årrække.

Svend Kreiner er professor emeritus ved *Center for Biostatistik, Københavns Universitet*. Svend Kreiner har forsket i test af elevers dygtighed siden slutningen af 60'erne. Han har publiceret en lang række forskningsartikler om test i internationalt anerkendte tidsskrifter og bøger. Han har gennem alle årene bidraget til udviklingen af tests inden for blandt andet matematik og læsning som bruges i skolen den dag i dag. I årene 2007-2010 bistod han Undervisningsministeriet ved udviklingen af nationale test, hvor han bl.a. udviklede nogle af de statistiske værktøjer som var nødvendige ved designet af en adaptiv test. Siden hen har han fungeret som sparringspartner for ministeriet ved spørgsmål om nationale test, og han har skrevet en række rapporter bl.a. om validering af testens resultater.

Jeppe Bundsgaard er professor MSO i fagdidaktik og it med særlig henblik på dansk ved *Danmarks institut for Pædagogik og Uddannelse (DPU), Aarhus Universitet*. Jeppe Bundsgaard har skrevet en række artikler om test og den pædagogiske brug af test fra et fagdidaktisk perspektiv. Han er dansk leder af den internationale IEA-undersøgelse *International Computer and Information Literacy Study* (ICILS 2013 og 2018), og han deltager som fagekspert nationalt og internationalt i PISA-undersøgelsen. Han har desuden deltaget i udviklingen af en række innovative computerbaserede test af elevers designkompetencer, samarbejdskompetencer mv.

Begge forfattere er udpeget til Undervisningsministeriets rådgivningsgruppe om evaluering af nationale test.

2.2 Rapportens opbygning

Denne rapport indeholder en omhyggelig analyse af de måletekniske egenskaber af de nationale test i læsning i 8. klasse i 2017. Analyserne bygger på svarene på alle opgaver fra alle 48.481 elever i 8. klasse i 2017. Analysen er foretaget ved hjælp af den samme Rasch-model som blev anvendt i forbindelse med udviklingen og afprøvningen af de nationale test. Eventuelle forskelle på det som analysen af datamaterialet fra 2017 måtte afsløre, i forhold til det som de nationale test producerer, kan derfor ikke skyldes at denne analyse bruger andre metoder og stiller andre krav til test, end dem der blev brugt i forbindelse med udviklingen af de nationale test.

Rapporten indledes i kapitel 3 med en præsentation af nationale test og af den statistiske model der ligger bag ved testen. I dette kapitel præsenteres også det anvendte datasæt. De to følgende kapitler (4 og 5) præsenterer analyserne og resultaterne af disse. I det sidste kapitel (kapitel 6) præsenteres de samlede konklusioner, og det diskuteres hvad konsekvenserne af problemerne er, og hvad der kan gøres for at rette op på disse problemer.

2.3 Resultaterne

Resultaterne af analysen kan ganske kort opsummeres på følgende måde, hvor det skal understreges at vores analyse kun kan fortælle noget om nationale tests målinger af elevers læsefærdighed i 8. klasse.

1) Nationale test anvender opgavernes sværhedsgrader til at udvælge opgaver til elever og til at beregne mål for hvor dygtige eleverne er. Disse sværhedsgrader er forkerte, og der var i 2017 mange tilfælde af meget store forskelle på de sværhedsgrader som nationale test benytter, og de sande sværhedsgrader.
2) Konsekvensen af at nationale tests sværhedsgrader er forkerte, er at beregningerne af dygtigheden er forkert og kan være direkte vildledende. Analyserne i kapitel 4 giver flere eksempler på at det rent faktisk er tilfældet, både når man ser på testresultater for enkelte elever, og når man ser på fordelingen af læsefærdigheden i 2017. Nationale tests resultater tegner med andre ord et forvrænget billede af situationen.
3) Anvendelsen af forkerte sværhedsgrader betyder at den adaptive algoritme vælger opgaver på en uhensigtsmæssig måde som – selvom målingerne af dygtigheden ikke havde været systematisk forkerte, og selvom analyserne også viser at den adaptive algoritme fungerer som den skal – ville forringe sikkerheden på målingerne. Resultaterne i kapitel 4 og 5 giver flere eksempler på at dette ikke blot er noget der kan påvises at følge logisk af problemerne med sværhedsgraderne, men at det også kan ses i data fra 2017.
4) Undersøgelsen i kapitel 4 af opgavernes tilpasning til eleverne viser at der i to af de tre områder som eleverne testes i (*profilområder*), kun er et meget begrænset udvalg af opgaver til dygtige og meget dygtige elever. Det kan betyde at disse elever vil opleve at få de samme opgaver i den obligatoriske test som de allerede har fået i den frivillige test, og det betyder at de skal besvare flere opgaver end nødvendigt for at opnå tilstrækkelig sikkerhed på resultatet.
5) Med hensyn til usikkerheden viser analyserne også at ministeriets beslutning om at slække på kravene til målingernes sikkerhed fører til testresultater der er så usikre at de er uanvendelige på elevniveau. Dette problem ville også være der selvom der ikke var problemer med opgavernes sværhedsgrader, men fejlene i sværhedsgraderne forstærker problemet med usikkerheden.
6) En undersøgelse af datas tilpasning til Rasch-modellen i kapitel 4 viser at der ikke er belæg i data for at sige at de tre profilområder i læsning måler én og samme færdighed, men analysen tyder på at det kan give mening at se afkodning og tekstforståelse som én dimension. En sådan sammenlægning af profilområder ville kunne forøge sikkerheden på elevernes resultater.
7) Analyserne i kapitel 5 viser at der er situationer hvor elevens færdigheder i læsning ikke kommer til udtryk i dele af testforløbet, og at beregningerne derfor systematisk undervurderer hvor godt eleven læser. De nævnte situationer, hvor dele af testforløbet er mislykkedes, er hyppigst i starten af forløbet, men forekommer også undervejs. Denne analyse kan ikke sige noget konkret om hvor ofte der er problemer med dele af testforløbene, men vi kan konstatere at det var særdeles let at finde sådanne eksempler ved blot at kigge datamaterialet igennem. Af den årsag forventer vi at hyppigheden af delvist mislykkede testforløb hvor elevernes færdigheder undervurderes, er med til at tegne et mere pessimistisk billede af danske elevers læsefærdigheder, end der er belæg for.

Kapitel 6 uddyber og diskuterer årsager til og konsekvenser af konklusionerne. I første omgang er det tilstrækkeligt at sige følgende med hensyn til konsekvenserne:

a) Problemerne med nationale tests målinger af læsefærdigheder i 8. klasse er så store at man nøje bør overveje om det er fagligt, pædagogisk og policy-mæssigt forsvarligt at fortsætte med at bruge denne del af nationale test før man har løst problemerne. Hvorvidt problemerne skal løses ved at erstatte nationale tests sværhedsgrader med sværhedsgrader baseret på aktuelle data, eller om der skal andre og mere drastiske løsninger til, kan analyserne i denne rapport ikke sige noget om.
b) Analysen i denne rapport er begrænset fordi den kun undersøger læsefærdighed i 8. klasse.

Forskellene på nationale tests resultater og resultaterne af 2017-analysen er imidlertid så slående at man skal være meget optimistisk hvis man forventer at lignende problemer ikke findes andre steder i nationale test. Af denne årsag anbefales det at man også nøje overvejer om det er hensigtsmæssigt at fortsætte med de øvrige områder af nationale test indtil man har undersøgt om tilsvarende problemer findes i disse, og indtil man har besluttet hvad der i givet fald skal gøres ved det.

c) Uanset hvad resultatet af overvejelserne bliver, skal lærere være opmærksomme på at testforløb kan være helt eller delvist mislykkede, at det kan lade sig gøre at undersøge det, og at man i givet fald bør forkaste testresultatet og undlade selv at bruge det samt undlade at orientere forældrene om andet end at testforløbet er slået fejl.

d) I forventning om at de sværhedsgrader som nationale test benytter, var korrekte i 2014, bliver konsekvensen af 2017-analysen at det ikke giver mening at sammenligne testresultater fra 2014 og 2017. At eventuelle forskydninger i fordelingerne af testresultaterne i form af ændrede gennemsnitsværdier eller ændringer i antallet af elever med gode eller dårlige testresultater ikke fortæller om læsefærdighederne har ændret sig og i givet fald på hvilken måde. Som følge heraf er det nødvendigt at gøre opmærksom på at testresultater fra nationale test ikke umiddelbart kan bruges til at vurdere om fordelingen af læsefærdigheder har ændret sig i positiv eller negativ retning som konsekvens af skolereformen.

3

De nationale test, måleegenskaber og Rasch-modellen

Dette kapitel introducerer kort til nationale test og beskriver dernæst dels de krav til skalaer som var vejledende i forbindelse med udviklingen af de nationale test i årene fra 2006 til 2009 og for vores analyser af datamaterialet fra 2017, dels det datamateriale som vi har benyttet i vores analyser.

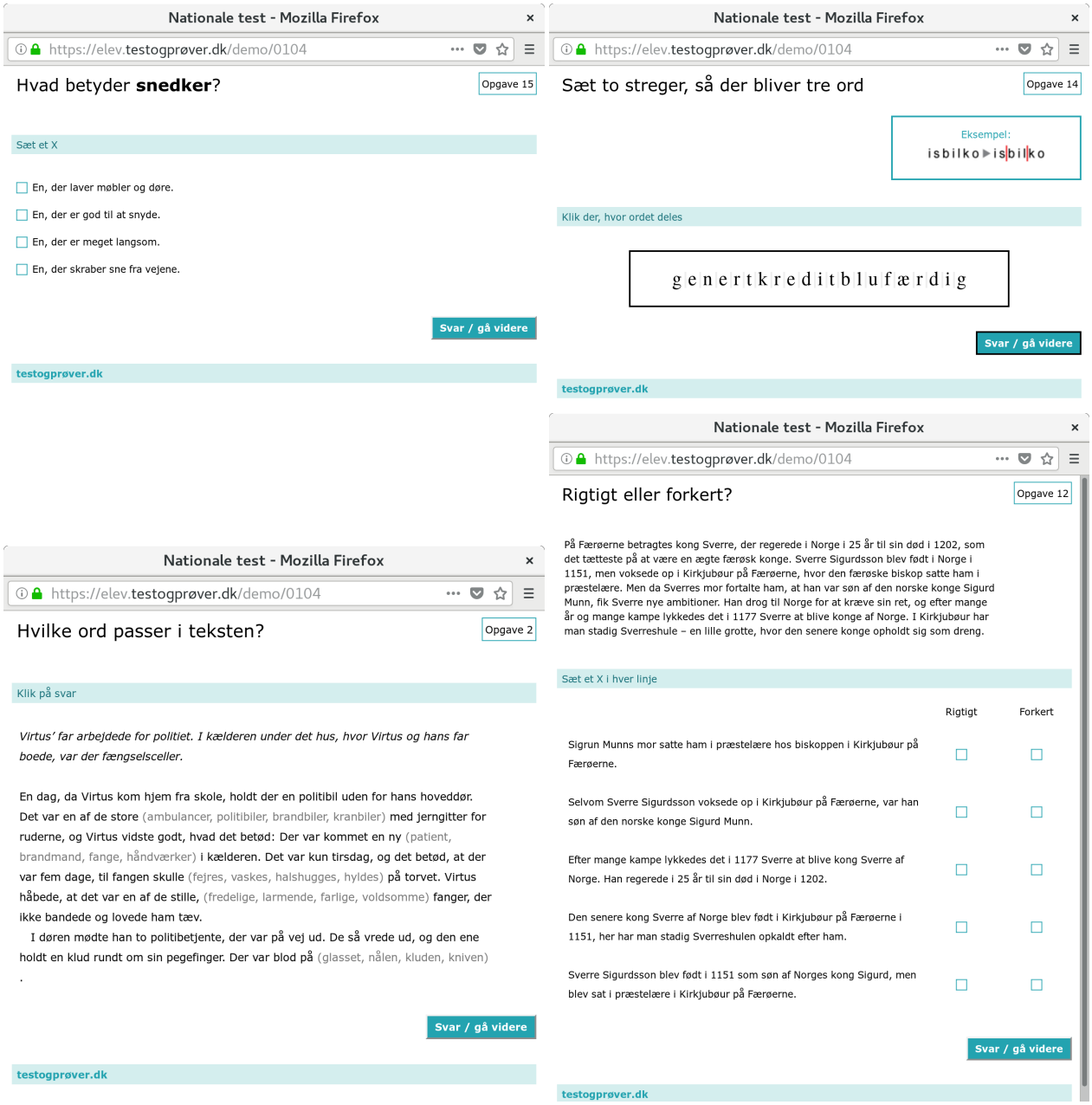
3.1 De nationale test

Elever i 2. til 8. klasse tager nationale test i fire forskellige fag eller fagområder fordelt over de syv årgange. Dansk, læsning tages i 2., 4., 6. og 8. klasse, og det tester som navnet siger, alene (dele af) læseområdet, men ikke de andre områder af danskfaget. Det er obligatorisk at deltage i nationale test for elever i folkeskolen. Ud over den obligatoriske test der finder sted i marts-maj i foråret, kan læreren eller skolen eller kommunen beslutte at eleverne skal tage testen "frivilligt" i efteråret før og i løbet af året efter den obligatoriske test. Testen er berammet til at tage 45 minutter. Men hvis ikke algoritmen er nået frem til en tilstrækkelig lav usikkerhed (SEM), kan læreren forlænge testforløbet.

I forbindelse med vedtagelsen af nationale test i 2006 blev det besluttet at man skulle måle tre såkaldte *profilområder* inden for hvert fagligt område. I dansk, læsning er de tre profilområder *sprogforståelse*, *afkodning* og *tekstforståelse*. Profilområderne svarer til det der i psykometrisk sprogbrug typisk omtales som *dimensioner*. Hvorvidt der i virkeligheden er tale om tre dimensioner der i psykometrisk forstand er forskellige, og hvorfor der netop skal være tre dimensioner i alle fagområder, er åbne spørgsmål som ikke er blevet besvaret, men som kan have uheldige konsekvenser for usikkerheden på målingerne.

Når man udvikler opgaver til en test, kan man vælge mange typer af opgaver. Det mest kendte er *multiple choice*, hvor testtageren får et spørgsmål og fx tre eller fire svar at vælge imellem (det kan fx være sætninger eller billeder)[1]. Det anvendes også i nationale test. Multiple choice-opgaver er blandt andet karakteriseret ved at det kan være vanskeligt at forudsige hvor svær opgaven er, fordi chancerne for at svare rigtigt afhænger af de forkerte svarmuligheder. Hvis de åbenlyst er forkerte, er der kun det rigtige svar tilbage, hvilket kan gøre et vanskeligt spørgsmål meget let. Desuden anvendes såkaldte *ordkædeopgaver*, hvor eleven skal sætte to streger i en kæde af bogstaver og derved danne tre ord. En tredje typisk opgave er såkaldte *cloze*-test (Taylor 1953) hvor eleven skal læse en tekst og undervejs angive hvilket ord der passer bedst i sammenhængen ud fra fx tre eller fire mulige. Eksempler på disse opgaveformater fra nationale test kan ses i figur 3.1.

[1] Faktisk er der tale om *single choince* fordi eleven skal vælge ét svar blandt flere mulige. Men i de fleste sammenhænge betegnes det *multiple choice*.



Figur 3.1 Eksempler på opgaveformater og opgaver i nationale test. Fra demo.testogprøver.dk.

Clozetests er kendetegnet ved det man kalder *lokal afhængighed*, dvs. at hvis man har et ord korrekt, er der større sandsynlighed for også at have de øvrige ord i samme tekst korrekte. Det er et problem i Rasch-modellen, og derfor kan man vælge at tælle antallet af rigtige delopgaver i stedet for at betragte hvert valg som én opgave. Sådanne opgaver kaldes ofte *superitems* eller *polytome* opgaver. Da man så ikke bare har items som kan være rigtige (1) og forkerte (0), men items som fx kan have værdierne 0, 1, 2, 3 osv., skal man bruge den nedenfor omtalte *partial credit-model*.

3.2 Krav til skalaer

De tekniske psykometriske[2] krav til gode test drejer sig dels om validitet og dels om usikkerhed og reliabilitet.

Rosenbaum (1989) opsummerer kravene til valide test på følgende måde:

a) Unidimensionalitet. Svarene på opgaver i en test må kun afhænge systematisk af én færdighed, der i princippet kan måles på en kvantitativ skala. På denne skala er en elev med en høj værdi dygtigere end en elev med en lav værdi, og forskellen på og/eller forholdet mellem de to tal kan betragtes som et udtryk for, hvor meget dygtigere den dygtigste elev er i forhold til den mindre dygtige.
b) Monotonicitet. For samptlige opgaver i testen skal sandsynligheden for et rigtigt svar på opgaven blive større og større, jo dygtigere eleven er.
c) Lokal uafhængighed. Sandsynligheden for at elev svarer rigtigt på en opgave, må ikke påvirke eller påvirkes af svarene på de øvrige opgaver.
d) Sandsynligheden for at en elev svarer rigtigt på en opgave, må ikke afhænge systematisk af andet end elevens dygtighed. Testen må med andre ord ikke indeholde opgaver der fx favoriserer drenge frem for piger, eller – i internationale undersøgelser – opgaver der favoriserer elever fra et land i forhold til andre.

Derudover var det i forbindelse med afprøvningen af nationale test et krav at opgavernes sværhedsgrader kan måles kvantitativt på den samme måde som det var tilfældet med dygtigheden[3]. Og det var et krav at opgaverne fungerer homogent, hvilket vil sige at forskellen på sværhedsgraden for to opgaver er den samme for de dygtigste elever og for de mindre dygtige elever.

Alle psykometriske skalamodeller er probabilistiske. De udtrykker sandsynligheden for at en elev kan svare rigtigt på en opgave, som en funktion af dygtigheden, sværhedsgraden og evt. andre forhold omkring opgaven og testsituationen. Af denne grund vil alle beregninger af hvor dygtige eleverne er, være statistiske estimater af dygtigheden som – ligesom alle andre statistiske estimater – vil være behæftet med en vis usikkerhed. I almindelige statistiske analyser måles graden af usikkerheden ved en såkaldt standardfejl (*standard error*, SE). Det er også tilfældet i forbindelse med pædagogiske test hvor standardfejlen omtales som *standard error of measurement* (SEM).

I alle statistiske analyser er det et krav at man forsøger at reducere usikkerheden i resultaterne ved at bruge metoder der reducerer standardfejlen til det mindst mulige. Det er også tilfældet i forbindelse med nationale test hvor man valgte et adaptivt testsystem fordi man kan vise at adaptive test resulterer i de mindst mulige SEM-værdier – vel at mærke hvis, og kun hvis, forudsætningerne for beregningen af dygtigheden og SEM er i orden.

I forbindelse med udviklingen af nationale test definerede man et krav om at SEM ikke måtte være større end 0.30. Baggrunden for denne beslutning vil ikke blive beskrevet her, selvom det ikke er uinteressant. Det viste sig nogle år efter testen var taget i brug, at dette mål ikke var nået, og at udviklerne af det adaptive system havde sat niveauet til 0,55 og ikke 0,3 (Ravn 2014; Wandall 2010; Styrelsen for It og Læring 2015). Konsekvenserne af denne beslutning vil blive diskuteret i forbindelse med vores konklusioner. På nuværende tidspunkt kan vi kun tage beslutningen til efterretning og vurdere om nationale test rent faktisk lever op til kravene.

Udover usikkerheden udtrykt ved SEM interesserer man sig i forbindelse med afprøvning af test også for et fænomen som i psykometrisk sprogbrug omtales som reliabilitet. Ordvalget er uheldigt fordi personer uden kendskab til psykometri opfatter begrebet som et generelt spørgsmål om hvor troværdige testresultaterne

[2] Psykometri er videnskaben om at måle (*-metri*) psykiske fænomener.

[3] Vi betegner elever der er i stand til at svare rigtigt på svære opgaver, som *dygtige* elever. Elever som kun kan svare på lettere opgaver, betegner vi som *mindre dygtige* elever. Betegnelsen gælder alene for det område eleven har svaret på opgaver inden for. Elever kan godt være dygtige inden for ét område og mindre dygtige inden for andre.

er (Styrelsen for It og Læring 2016b). For at forhindre misforståelser er det derfor nødvendigt at definere reliabiliteten.

Reliabiliteten afhænger af usikkerheden (SEM), men den afhænger i lige så høj grad af spredningen af eleverne i den population man interesserer sig for. Definitionen er:

$$\text{Reliabilitet} = \frac{\text{Variansen af dygtigheden}}{\text{Variansen af dygtigheden} + SEM^2}$$

Det vil sige at reliabiliteten forbedres jo mere elevernes scorer spreder sig, og forværres jo mindre spredningen er. Hvis alle elever er lige dygtige, vil reliabiliteten altid være lig med nul selvom testen i sig selv er meget sikker.

Reliabilitetsmålet er en abstrakt størrelse som kan være vanskelig at fortolke. Den bedste måde at forstå den på er ved at bemærke at sandsynligheden for at den dygtigste af to tilfældigt udvalgte elever vil få det bedste testresultat, vil være tæt på (men ikke helt præcist lig med) reliabiliteten. Hvis reliabiliten fx er 0,8, vil den bedste af to tilfældigt udvalgte elever blive vurderet som den bedste i otte ud af ti målinger og som den dårligste i to af disse ti målinger.

I forbindelse med test der skal anvendes på individniveau, finder man ofte et krav om et reliabiliteten skal være lig med 0,9 eller bedre (Hale og Astolfi 2014; Wells og Wollack 2003). I populationsanalyser (fx PISA og IEA-undersøgelserne), hvor det ikke drejer sig om enkelte elever, er kravet at reliabiliteten skal være større end 0,7 og helst større end 0,8.

Baggrunden for disse krav vil ikke blive diskuteret i denne rapport hvor formålet først og fremmest er at undersøge om nationale test leverer lige så gode og pålidelige testresultater som man lover.

3.3 Rasch-modellen

Georg Rasch var matematiker og arbejdede blandt andet ved Danmarks Pædagogiske Institut i 1950-60'erne. I den periode udviklede han det der i dag er kendt som Rasch-modellen (Rasch 1960).

I virkeligheden er Rasch-modellen en hel familie af item-responsmodeller som deler en række specielle og særdeles nyttige statistiske egenskaber.

Den oprindelige Rasch-model blev første gang omtalt på dansk i Rasch (1958) og senere på engelsk i Rasch (1960). Modellen beskriver sandsynligheden for at en elev svarer rigtigt på en opgave i en pædagogiske test som en funktion af forskellen på elevens dygtighed, θ (theta), og opgavens sværhedsgrad, β (beta):

$$\text{Sandsynlighed for korrekt svar} = \frac{e^{\theta-\beta}}{1+e^{\theta-\beta}}$$

Bortset fra at eksponentialfunktionen optræder i formlen, er der tale om en i princippet meget enkel model der blot siger det intuitivt indlysende at en elevs sandsynlighed for at svare rigtigt på en opgave afhænger af to ting (og kun disse to ting), nemlig elevens dygtighed (der ofte benævnes med det græske bogstav theta, θ) og opgavens (*itemmets*) sværhedsgrad (der ofte benævnes beta, β).

Ifølge denne model er sandsynligheden for et korrekt svar lig med 50 procent hvis dygtigheden er lig med sværhedsgraden: Eleven får altså ikke en let opgave, hun får en udfordrende opgave som hun har en fair chance for at kunne besvare.

Raschs oprindelige model, hvor der kun er to mulige udfald, rigtigt eller forkert, omtales i dag som Rasch-modellen for dikotome items. Denne model er senere, først af Rasch selv i 1961 og 1962 og senere af andre udvidet til modeller for polytome items hvor der kan scores fra 0 til et vist antal point afhængig af hvor

stor en del af opgaven der er besvaret rigtigt. Formlerne for denne model, der som regel omtales som en *partial credit model* (PCM), er mere komplicerede. Hvis der kan scores fra 0 til m point på opgaven, er sandsynligheden for at der scores x point givet ved:

$$\text{Sandsynlighed for x point} = \frac{e^{\sum_{j=0}^{x} \theta - \beta_j}}{\sum_{z=0}^{m} e^{\sum_{j=0}^{z} \theta - \beta_j}}$$

hvor m er det maksimale antal point på itemmet.

I PCM-modellen omtales β-parametrene normalt som "tærskelværdier". Læseren er undskyldt hvis hun har vanskeligt ved at gennemskue præcis hvilken fortolkning disse parametre har. Det har alle andre også. Det der er værd at bemærke, er at sandsynlighederne defineres af forskellene på dygtigheden og tærskelværdierne, og at det forventede antal point er en monotont voksende funktion af dygtigheden. Jo dygtige, jo flere point på opgaven.

Rasch-modellen indeholder ingen forudsætninger om at dygtigheden fordeler sig på en bestemt måde i en population af elever. Sådanne forudsætninger optræder til gengæld i den generalisering af Rasch-modellen der omtales som *Mixed-Coefficient Multinomial Logit Model* (MCMLM) eller *Multidimensional Random Coefficients Multinomial Logit Model* (MRCMLM) (Adams, Wilson, og Wang 1997), og som blandt andet har været anvendt i analyserne i denne rapport.

Disse modeller har den styrke at de tilbyder en enkel og naturlig generalisering af den endimensionelle Rasch-model der kun måler én færdighed ad gangen, til én samlet model for flere test der hver for sig måler forskellige færdigheder.

Da nationale test ikke måler én færdighed, men derimod tre kvalitativt forskellige færdigheder som i nationale tests private terminologi omtales som profilområder, er MCMLM-modellen særdeles relevant at anvende. For testen i dansk, læsning er disse profilområder *sprogforståelse*, *afkodning* og *tekstforståelse*.

En MCML-model for nationale test måler altså dygtigheden for tre forskellige færdigheder, (θ_1, θ_2, θ_3). Modellen bygger på følgende forudsætninger:

a) At opgaver knyttet til et bestemt profilområde kun afhænger af dygtigheden knyttet til profilområdet.
b) At opgaverne til et bestemt profilområde isoleret set opfører sig som opgaver fra en endimensionel Rasch-model.
c) At fordelingen af de tre dygtighedsgrader (θ_1, θ_2, θ_3) følger en flerdimensionel normalfordeling hvor sammenhængen mellem de tre færdigheder er lineær og kan beskrives fyldestgørende ved korrelationen mellem dygtighedsgraderne.

MCML-modellen er således blot en flerdimensionel Rasch-model.

Den statistiske analyse estimerer modellens parametre og afprøver om data (dvs. svarene på opgaverne) opfører sig som Rasch-modellen forventer. Detaljerne i metoderne vil ikke blive gennemgået her, men der er en ting som skal understreges.

Estimatet af en elevs dygtighed inden for et bestemt profilområde er en relativt kompliceret ikke-lineær funktion af det samlede antal korrekte opgaver eller det samlede antal point eleven har opnået, som ikke – fordi det er en Rasch-model og ikke en af de andre skalamodeller – afhænger af hvordan eleven har scoret pointene. Hvis tre elever hver for sig har seks rigtige ud af ti opgaver, vil de få det samme estimat af dygtigheden selvom svarmønstret for den ene var (0,0,0,0,1,1,1,1,1,1), for den anden var (0,1,1,0,1,0,1,1,1,0) og for den tredje var (1,1,1,1,1,1,0,0,0,0). Det første og det sidste mønster ser måske mærkelige ud, men hvis der fx var tale om ti lige vanskelige opgaver ville de tre forløb faktisk have præcis samme sandsynlighed for at forekomme.

Eller med andre ord: Hvis opgaverne virkelig opfører sig som Rasch-modellen kræver – og kun hvis det er tilfældet – er det ligegyldigt hvorledes pointene er scoret.

Tabel 3.1 Oversigt over den bedst mulige "standard error of measurement" (SEM) som funktion af antal besvarede opgaver

Antal opgaver	Mindste mulige SEM
1	2,00
3	1,15
5	0,89
10	0,63
20	0,45
30	0,37
40	0,32
50	0,28
60	0,26
70	0,24

3.4 Adaptive test

De nationale test er som nævnt adaptive og benytter en adaptiv algoritme der løbende beregner estimater af dygtigheden ud fra oplysninger om sværhedsgraderne for de opgaver der allerede er besvaret, og antallet af disse opgaver der er besvaret korrekt, hvorefter der vælges en ny opgave med en sværhedsgrad der ligger tæt på det aktuelle bud på dygtigheden (Styrelsen for It og Læring 2015). Interessen for adaptive test udspringer af bekymringen for usikkerheden i testresultaterne. Usikkerheden måles ved hjælp af standardfejlen (SEM) på estimatet af dygtigheden. SEM kan vises at være (næsten) lig med 1 divideret med kvadratroden på den såkaldte test-information.

Test informationen er på sin side defineret som summen af den information som svaret på en enkelt opgave bidrager med. Denne såkaldte iteminformation har to forskellige egenskaber. For det første kan den vises at være størst og at være lig med 0,25 (for dikotome items) hvis opgavens sværhedsgrad er lig med elevens dygtighed, og for det andet kan den vises at blive mindre og mindre jo større forskel der er på dygtigheden og sværhedsgraden. Det svarer til at man ikke kan forvente at få noget ud af at stille en alt for let opgave til en dygtig elev. Man ved jo på forhånd at man vil få et korrekt svar. Af disse egenskaber følger det at testinformationen er størst hvis samtlige opgaver har sværhedsgrader der ligger tæt på dygtigheden, og at den aldrig kan blive større end 0,25 ganget med antallet af besvarede opgaver. Eller med andre ord: SEM kan aldrig kan blive mindre end 2 divideret med kvadratroden af antallet af besvarede opgaver.

I tabel 3.1 ses en række af de værdier man kan opnå under perfekte forhold. Tabellen viser at hvis man vil ned på en SEM på 0,30 for et bestemt profilområde, således som det oprindelig blev formuleret for nationale test, så skal der stilles mellem 40 og 50 opgaver inden for profilområdet, og disse opgaver skal have sværhedsgrader der ikke ligger alt for langt fra elevens dygtighed.

Det er disse egenskaber der er baggrunden for valget af adaptive test. Nationale test er tænkt som et redskab, der skal give ikke alt for usikre målinger for alle elever uanset. Hvis det skal opnås, er man nødt til at stille forskellige opgaver til forskellige elever. De dygtigste elever skal have vanskelige opgaver, og de mindst dygtige skal have lette opgaver. Og den teknisk bedste måde at sikre det på er ved at anvende adaptive test[4].

Bemærk, at der er mest information ved opgaver som eleverne har 50 procents sandsynlighed for at svare rigtigt på. Der er således tale om opgaver som er udfordrende for eleven, men som de har en fair chance for at besvare. En adaptiv test som forsøger at optimere informationen, vil altså stille opgaver som udfordrer

[4]Det skal understreges at vi ikke dermed har sagt at en adaptiv test nødvendigvis er den bedste og billigste måde at teste elever på. En adaptiv test stiller nemlig en række krav til opgaverne og nødvendiggør et meget dyrt udviklingsarbejde som fjerner midler fra andre aspekter af udviklingen, fx udvikling af innovative testopgaver.

eleven maksimalt. Og derfor vil eleverne i sagens natur opleve at testen er svær, og at de skal tænke over hvert eneste spørgsmål.

Den adaptive algoritme er i princippet særdeles enkel idet den består af tre trin:

1) Processen indledes med et mindre antal opgaver (i nationale test drejer det sig om tre opgaver), hvorefter det første estimat af dygtigheden beregnes.
2) Derefter vælges en enkelt opgave med en sværhedsgrad der ligger tæt på det aktuelle estimat af elevens dygtighed. Der er intet krav om at forskellige elever skal have forskellige opgaver, og heller ikke noget krav om at alle opgaver skal bruges. Kravet er kun at opgaven passer til Rasch-modellen, at sværhedsgraderne er korrekte, og at der er tilstrækkelig mange opgaver til alle elever fra de mindst dygtige til de dygtigste.
3) Det aktuelle estimat opdateres, og SEM genberegnes. Hvis SEM er mindre end kravet til SEM, eller hvis tiden er udløbet, stopper algoritmen. I modsat fald vender systemet tilbage til trin 2.

Usikkerheden målt ved SEM vil være betragtelig i starten af forløbet. Det vil især være tilfældet hvis algoritmen vælger opgaver der åbenlyst er enten for lette eller for vanskelige for eleven, som det er tilfældet i nationale test. Hvis man plotter det aktuelle estimat af dygtigheden, kan der være meget store variationer i starten, og udgangspunktet kan ligge et helt andet sted end der hvor algoritmen stopper. SEM vil jo være større end 1,15 og ofte meget større efter de første tre opgaver. Det er uundgåeligt og ikke et udtryk for at der er noget galt med opgaver eller den adaptive algoritme.

Bemærk også, at estimatet genberegnes efter hver skridt på en måde der kun afhænger af hvor mange point der er scoret. Det foregående estimat indgår ikke i beregningerne. Det er altså ikke et spørgsmål om at man ser på hvordan det foregående estimat så ud, og derefter korrigerer for at tage hensyn til det nye estimat. Forløbet (0,0,0,1), hvor 0 er forkert svar, og 1 er rigtigt, vil derfor give samme estimat efter fjerde trin som forløbet (0,1,0,0).

Hvis forudsætningerne i form af tilpasning til Rasch-modellen er opfyldte, vil en sådan adaptiv algoritme fungere så godt som muligt. Hvis forudsætningerne ikke er opfyldt, vil den adaptive algoritme give forkerte resultater og større usikkerhed der gør forskellige testresultater usammenlignelige, både når det drejer sig om sammenligninger mellem forskellige elever, sammenligninger mellem forskellige elevpopulationer og sammenligninger mellem flere parallelle testresultater for den samme elev. Testresultaterne vil være uanvendelige og vildledende. Helt det samme vil ikke nødvendigvis være tilfældet i en lineær test hvor alle får samme opgaver, for her vil alle blive udsat for de samme systematiske fejl. Resultatet vil altså være lige forkert for alle.

Årsagen til problemerne er enkle:

1) Hvis sværhedsgraderne er forkerte, vælger systemet forkerte opgaver der ikke passer godt nok til eleverne. Det vil forøge usikkerheden (SEM).
2) Hvis sværhedsgraderne er forkerte, ville der være tale om systematiske fejl hvis det var en lineær test. Man ville ikke kunne se disse ved gentagen brug af testen for den samme elev. Men når der er tale om adaptive test, vil eleverne blive præsenteret for forskellige opgaver med hver sine fejl. Det vil få testresultater for eleven til at variere endnu mere end de skulle på grund af den almindelige usikkerhed som testresultater er behæftet med. Og det vil udsætte forskellige elever for forskellige fejl.
3) Hvis der derudover også er problemer med tilpasningen til Rasch-modellen, vil systemet bruge forkerte formler til beregning af dygtigheden hvilket vil bidrage yderligere i forhold til de fejl der blev nævnt under punkt 2.

Der er mange rapporter om påfaldende stor variation ved sammenligning af de samme elevers resultater ved obligatoriske og frivillige nationale test (Ravn 2015c; Norling 2016). Og der er mange beskrivelser af testforløb der ved en umiddelbar betragtning ser meget mærkelige ud. Det var for at undersøge og eventuelt afkræfte en hypotese om at disse iagttagelser kan skyldes den form for modelfejl som omtales ovenfor, at analyserne i denne rapport er foretaget.

Tabel 3.2 Uddrag af det originale datasæt

elev_id	fag	profilomraade	svartidspunkt	opgavenummer	score	no_items	theta
100008	Dansk/læsning 8. klasse	1	02FEB2017:08:01:31.823	0108010000301238607-1	0	1	0
100008	Dansk/læsning 8. klasse	2	02FEB2017:08:02:09.863	0108020111018	1	1	2.5

Note:
Observationerne angiver elev-id, fag, profilområde, tidspunkt for besvarelse, hvilken opgave der er besvaret, hvor mange delspørgsmål eleven havde rigtige (score), hvor mange delspørgsmål (items) opgaven bestod af, og hvad den estimerede dygtighed var efter besvarelsen af spørgsmålet (theta). For hver elev er der typisk i omegnen af 50-150 poster.

Tabel 3.3 Uddrag af det berigede datasæt

elev_id	fag	profilomraade	svartidspunkt	opgavenummer	score	no_items	theta	responseNo	proResponseNo	timeD	theta2017	SEM2017	betaDNT	beta2017
100008	Dansk/læsning 8. klasse	1	02FEB2017:08:01:31.823	0108010000301238607-1	0	1	0,0	1	1		-0,438	2,16	0,591	0,982
100008	Dansk/læsning 8. klasse	2	02FEB2017:08:02:09.863	0108020111018	1	1	2,5	2	1	0,634	1,817	2,50	2,069	1,017

Note:
De tilføjede variable er:
[1] responseNo: opgavens nummer i den rækkefølge eleven har modtaget dem.
[2] proResponseNo: opgavens nummer inden for det givne profilområde.
[3] timeD: tid brugt på opgaven (beregnet som forskel i svartidspunkt, og derfor ikke opgjort for den første opgave).
[4] theta2017: estimeret dygtighed efter besvarelsen af spørgsmålet ifølge analysen foretaget i forbindelse med denne rapport.
[5] SEM2017: usikkerheden på resultatet efter besvarelsen af spørgsmålet.
[6] betaDNT: opgavens sværhedsgrad ifølge nationale tests Rasch-analyse.
[7] beta2017: opgavens sværhedsgrad ifølge denne rapports Rasch-analyse.

3.5 Karakteristika ved det anvendte datasæt

Denne rapport er som sagt baseret på data fra nationale test i dansk, læsning i 8. klasse fra foråret 2017. Konkret består datasættet i en observation for hver elevbesvarelse af et spørgsmål (se tabel 3.2). 48.481 elever gav samlet 2.747.092 svar på 823 opgaver i nationale test i dansk, læsning i 2017.

Til brug for analyserne har vi gennemført en række beregninger af nye variable som vi har koblet til det oprindelige datasæt. Disse kan ses i tabel 3.3.

Vi har desuden udarbejdet en række omformede datasæt til brug for en Rasch-analyse (se kapitel 4) og for en undersøgelse af mønstre i besvarelser (se kapitel 5).

I tabel 3.4 ses en række oplysninger om nationale test som er beregnet ud fra datasættet.

Som det fremgår, har nogle elever formået at tømme itembanken fuldstændigt, altså svaret på alle de tilgængelige opgaver. Der er dog tale om meget få elever, og der kan være tale om at de blot har klikket sig igennem uden at svare. 134 gange har en elev besvaret 100 opgaver eller mere inden for et profilområde; i 86 af disse forløb har den samme elev besvaret mere end 100 opgaver inden for to eller tre profilområder (altså samlet mere end 200 eller 300 opgaver). Gennemsnittet af items per profilområde er 20-22 items, svarende til et samlet gennemsnit per elev på 62 items per test (der kan være flere items per opgave i profilområde 3).

I profilområde 2 er der en opgave som mere end tre femtedele af eleverne ser, og i profilområde 3 er der en opgave som mere end en tredjedel af eleverne ser. I gennemsnit ses en opgave af mellem fem til ti procent af eleverne.

En test tager i gennemsnit 53 minutter. For 50 procent tog den mere end 47 minutter, for 25 procent varede den mere end 59 minutter, og for ti procent varede den mere end 72 minutter. For de knap 500 elever (én procent) der sad længst ved testen, varede den mere end 100 minutter.

Tabel 3.4 Centrale informationer om nationale test, 8. klassse, dansk, læsning

			Items per elev		Elever per item			Tidsforbrug			
Profilområde	Opgaver	Items	Gns.	Maks	Min	Gns.	Max	Opgave-gns.	Test-gns.	Test-median	Top 1% på Test
1	308	308	19,5	308	837	3065	6629	0,5	52,9	47,3	101
2	214	214	19,8	214	157	4479	30216	0,6			
3	301	634	22,2	587	260	2806	18438	1,9			

Note:
Varigheden af tid på opgaverne er beregnet ved at trække svartidspunktet for den foregående opgave fra denne opgaves svartidspunkt. Den første opgave i et forløb har derfor ikke nogen tid. Ved beregningen af gennemsnitlig (forkortet gns.) tid brugt på opgaverne og af samlet tid på testen er alle tider over 10 minutter betragtet som ukendte (der har sandsynligvis været holdt pause), og de indgår derfor ikke i beregningen af gennemsnit osv.

4

Item-analyser af data fra 2017

4.1 Indledning

Formålet med de følgende analyser er at undersøge om de historiske sværhedsgrader som nationale tests adaptive algoritme benytter sig af, svarer til de aktuelle sværhedsgrader estimeret i data fra 2017.

For at undersøge nationale tests måleegenskaber er der brug for en sammenligning med noget der anses for korrekt. For nationale test er itemsværhedsgraderne oprindelig estimeret i forbindelse med den første gennemførsel af testen i 2010. I 2014 foretog ministeriet ifølge en artikel i Folkeskolen en fornyet Rasch-analyse af itembankerne (Ravn 2015a)[1]. Når eleverne tager testen, anvendes de på forhånd estimerede itemsværhedsgrader til at estimere elevernes dygtigheder. Rasch-modellen bygger på en empirisk underbygget hypotese om at elever der tager de samme opgaver, vil opleve dem som tilsvarende svære. Om dette er tilfældet kan testes ved at genestimere opgavernes sværhedsgrader og elevernes dygtigheder. Det er en sådan analyse denne rapport præsenterer.

Udgangspunktet for sammenligningerne er at estimaterne af sværhedsgraderne i 2017 er et udtryk for opgavernes sande sværhedsgrader for de elever der tog testen i 2017. Der er naturligvis en vis usikkerhed forbundet med estimaterne, men datamaterialet er som omtalt i kapitel 3, særdeles omfattende således at usikkerheden af estimaterne er begrænset. Hvis nationale tests sværhedsgrader passer til virkeligheden som den ser ud i 2017, skal forskellen på nationale tests sværhedsgrader og 2017-estimaterne være meget begrænsede.

Hvis der er god overensstemmelse mellem de to sæt sværhedsgrader, skal der også være en tilsvarende god overensstemmelse mellem estimaterne af færdigheden og de beregnede SEM-mål for målingernes usikkerhed. Hvis de to sæt sværhedsgrader er forskellige, vil målingerne af færdighederne også være forskellige, og usikkerheden vil være større i de sande 2017-estimater af færdighederne end i de forkerte nationale test-estimater fordi den adaptive rutine vælger forkerte opgaver der bidrager med mindre information end hvis systemet havde valgt nye opgaver baseret på de korrekte sværhedsgrader. Disse forhold vil blive beskrevet hvis sammenligningen mellem de to sæt sværhedsgrader falder negativt ud.

4.2 Statistiske metoder

Den Rasch-model der anvendes til analyserne af data fra 2017, er den samme som den der blev anvendt i forbindelse med den oprindelige afprøvning af nationale test. Forudsætningerne for analysen af data fra 2017 er derfor de samme som de forudsætningerne, der lå bag den oprindelige afprøvning af nationale test, og der er ikke tilføjet nye og mere restriktive forudsætninger i 2017-analysen. Det er med andre ord muligt at afprøve kvaliteten af nationale test i 2017 på nationale tests egne præmisser.

Med hensyn til de statistiske metoder der anvendes, er der små forskelle på analyserne i forbindelse

[1] Vi har ikke været i stand til at finde dokumentation fra Undervisningsministeriet der beskriver denne reestimation af itemsværhedsgraderne.

med afprøvningen og analyserne af data fra 2017 fordi vi har anvendt metoder til at estimere item- og personparametre der er lidt bedre end dem der blev anvendt i forbindelse med de tidligere analyser. Da der i alle tilfælde er tale om analyser af meget store datamaterialer vil forskellen på de forskellige metoder til at estimere Rasch-modellens parametre være uden praktisk betydning. Den usikkerhed som alle former for statistiske beregninger er behæftet med, vil for begge metoder være meget begrænset på grund af datamaterialernes størrelser. Vi betragter stadig de metoder man tidligere benyttede, som gode nok, og man kan uden problemer tillade sig at sammenligne estimaterne fra 2010 og 2014 med estimaterne fra 2017. Hvis der er store forskelle, er det fordi sværhedsgraderne i 2017 afviger fra de tidlige sværhedsgrader. Det vil ikke være en konsekvens af at der bruges forskellige metoder til at estimere dem eller et resultat af den statistiske usikkerhed som estimaterne er behæftet med.

4.2.1 Beregning af tal for sværhedsgrader og dygtighed

Rasch-modellens item- og personparametre der er udtryk for henholdsvis opgavernes sværhedsgrader og elevernes dygtighed, estimeres for hvert profilområde for sig.

Estimation af parametre i Rasch-modellen foregår i to trin, hvor det første trin estimerer opgavernes sværhedsgrader, og hvor det andet trin bruger disse estimater til at beregne estimater af dygtigheden for de enkelte elever. De nationale tests sværhedsgrader er estimeret ved hjælp af såkaldte *pairwise estimates* (Zwinderman 1995) der har den fordel at estimaterne ikke stiller krav om at personerne er fordelt på en bestemt måde. Da pairwise estimates ikke udnytter det indsamlede datamateriale fuldt ud, er disse estimater behæftet med en lidt større usikkerhed end andre estimater af Rasch-modellens item-parametre.

Opgavernes sværhedsgrader i 2017 estimeres vha. såkaldte *marginal estimates* (Boch og Aitken 1981). Disse estimater stiller ingen krav til dygtigheden for de enkelte elever, men det antages at fordelingen af dygtigheden er tilnærmelsesvis normalfordelt, og middelværdien og spredningen i denne fordeling estimeres samtidig med at sværhedsgraderne estimeres. De marginale estimater er naturligvis følsomme over for meget store afvigelser fra antagelse om normalfordelingen, men estimaterne udnytter til gengæld data fuldt ud og vil derfor være mindre usikre end andre estimater af Rasch modellens sværhedsgrader. Som det vil fremgå af det følgende, er forudsætningerne for anvendelsen af marginale estimater opfyldt.

I forbindelse med analysen af data fra 2017 beregnes tal for dygtigheden ved hjælp af såkaldte *Weighted Maximum Likelihood* (WML) estimater, mens nationale test beregner dygtigheden ved hjælp af almindelige *Maximum likelihood*-metoder (ML-metoder). Der henvises til Kreiner og Christensen (2013) for en nærmere forklaring af forskellen på de to estimater. Alle estimater af dygtigheden i Rasch-modellen er præget af en vis grad af systematisk bias for elever der har besvaret mange opgaver der enten er for lette for eleven eller for vanskelige for eleven. Valget af WML-estimater i denne analyse skyldes at de er præget af mindre systematisk fejl end andre estimater. Brugen af ML-estimater i nationale test skyldes at adaptive test kontrollerer at der hverken er for mange lette eller for mange vanskelige opgaver for eleverne, således at det kan forudses at nationale tests estimater af dygtigheden ikke vil have de samme problemer med systematiske fejl som ML-estimater kan have i almindelige test. I forbindelse med analyserne af enkelte testforløb i kapitel 5 og bilag B er der beregnet både WML- og ML-estimater for at undersøge om forventningerne til ML-estimaterne er opfyldt. Det viser sig at være tilfældet.

Da estimeringen af dygtigheden udnytter værdierne af sværhedsgraderne, følger det at dygtigheden altid bestemmes relativt til opgavernes sværhedsgrader, og at tallene for dygtigheden ikke kan betragtes som absolutte mål for dygtigheden. Forskellen på dygtighederne for to elever fortæller hvor meget dygtigere den ene er i forhold til den anden, men tallene i sig selv siger intet om hvorvidt den dygtigste i faglig forstand er dygtig og om den mindst dygtige er udygtig.

Udover det fortæller teorien for Rasch-modeller at tallene for sværhedsgraderne og tallene for dygtigheden kan placeres på en og samme intervalskala der i Rasch-modellens terminologi omtales som en logit-skala[2]. Det vil føre for vidt her at forklare hvorfor man anvender denne terminologi, og det er i øvrigt uden praktisk

[2]Logits er tal der fremkommer ved følgende transformation af sandsynligheder målt på skalaer fra 0 til 1. Hvis p er lig med sandsynligheden, er den tilsvarende logit-værdi lig med den naturlige logaritme af p/(1-p).

Tabel 4.1 Centrale parametre fra analysen af data fra nationale test.

Analyse	Antal Items	Elever	Reliabilitet	Varians	Antal Parametre
Profilområde 1	308	48481	0,796	1,031	309
Profilområde 2	214	48481	0,778	0,904	215
Profilområde 3	301	48481	0,779	0,859	635

[1] Reliabilitet: Reliabilitetsmål baseret på TAM-pakkens Estimated A Posteriori-estimat af elevernes dygtigheder.
[2] Varians: Varians af elevparametrene.
[3] Antal parametre: Antal parametre som estimationen har udregnet (items plus steps plus gennemsnit).

betydning. I denne sammenhæng er det tilstrækkeligt at gøre opmærksom på at denne rapport bruger samme terminologi fordi den vil være genkendelig for eksperter med forstand på Rasch-analyser.

4.2.2 En, to eller tre latente dimensioner

Nationale test antager at de tre profilområder måler tre kvalitativt – men naturligvis korrelerede – forskellige latente dimensioner. Tre forskellige færdigheder. Konsekvensen af denne antagelse er at der bliver relativt lidt tid til hvert profilområde og derfor forholdsvis få opgaver og stor usikkerhed per profilområde. Valget af netop tre profilområder i hvert eneste fag har – så vidt vi ved – aldrig været fagligt motiveret, og minder mere om en administrativ beslutning som er truffet for at gøre udviklingen af nationale test så enkel som mulig.

Da beslutningen har konsekvenser for sikkerheden, og da nationale test har været kritiseret for at give for usikre resultater, har vi i denne rapport inkluderet ekstra analyser der kan belyse om nationale tests profilområder måler en, to eller tre forskellige dimensioner. Denne analyse er foretaget ved hjælp af marginale metoder til analyse af multivariate Rasch-modeller (Adams og Wu 2007) der antager at de forskellige færdigheder der ligger bag de tre profilområder, fordeler sig som multivariate normalfordelinger, og som estimerer korrelationer og kovarianser mellem profilområderne samtidig med estimationen af middelværdier og varianser.

Alle analyser i dette kapitel er foretaget ved hjælp af R-pakken TAM. Der henvises til dokumentationen for dette program (Robitzsch, Kiefer, og Wu 2019) og til dokumentationen for ConQuest (Wu m.fl. 2007) som TAM bygger på, samt Adams og Wilson (1996) og Adams, Wilson, og Wang (1997) for yderligere oplysninger om multivariate Rasch-modeller og itemanalyse ved hjælp af marginale metoder. Scriptet der er anvendt, kan rekvireres hos forfatterne (tillige med alle de øvrige scripts der er udviklet til denne rapport).

Analysen er foretaget på datasættet bestående af alle elevbesvarelser af nationale test i dansk, læsning for 8. klasse i 2017. Vi kalder i det følgende denne analyse for *2017-analysen*, mens analysen som nationale test viser på, kaldes *DNT-analysen*. Til brug for denne analyse er datasættet omkodet så der er en række for hver af de 48.481 elever og en kolonne for hver af de 823 opgaver. Derved er hver elevs score (rigtigt/forkert) registreret for hver opgave. Centrale parametre fra analysen er gengivet i tabel 4.1

4.3 Sammenligning af sværhedsgrader og elevdygtigheder i 2017-analysen og nationale tests analyse

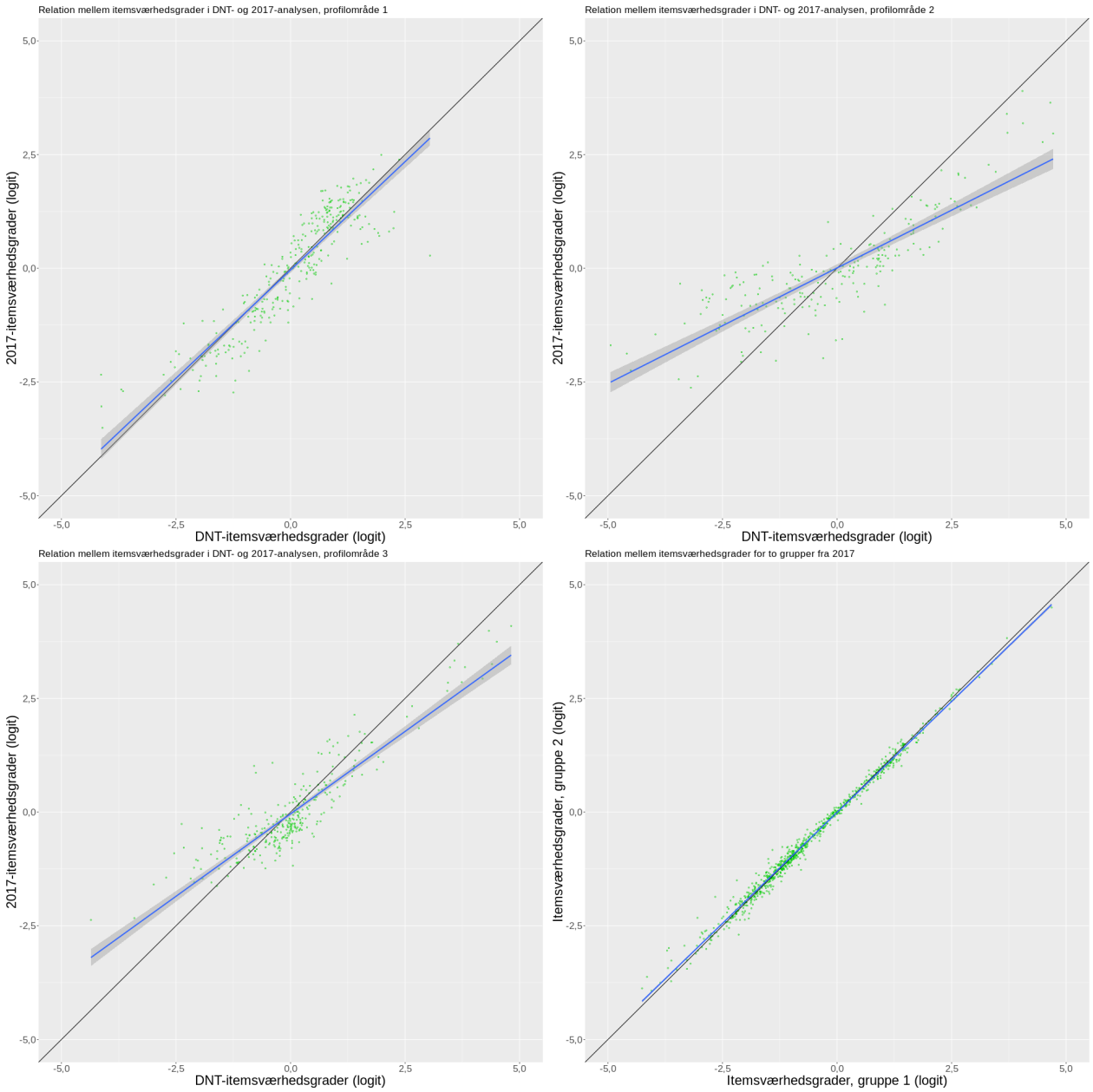
4.3.1 Sammenligning af sværhedsgrader

Bilag C indeholder en tabel over sværhedsgraderne ifølge DNT og sværhedsgraderne ifølge analysen af 2017-materialet.

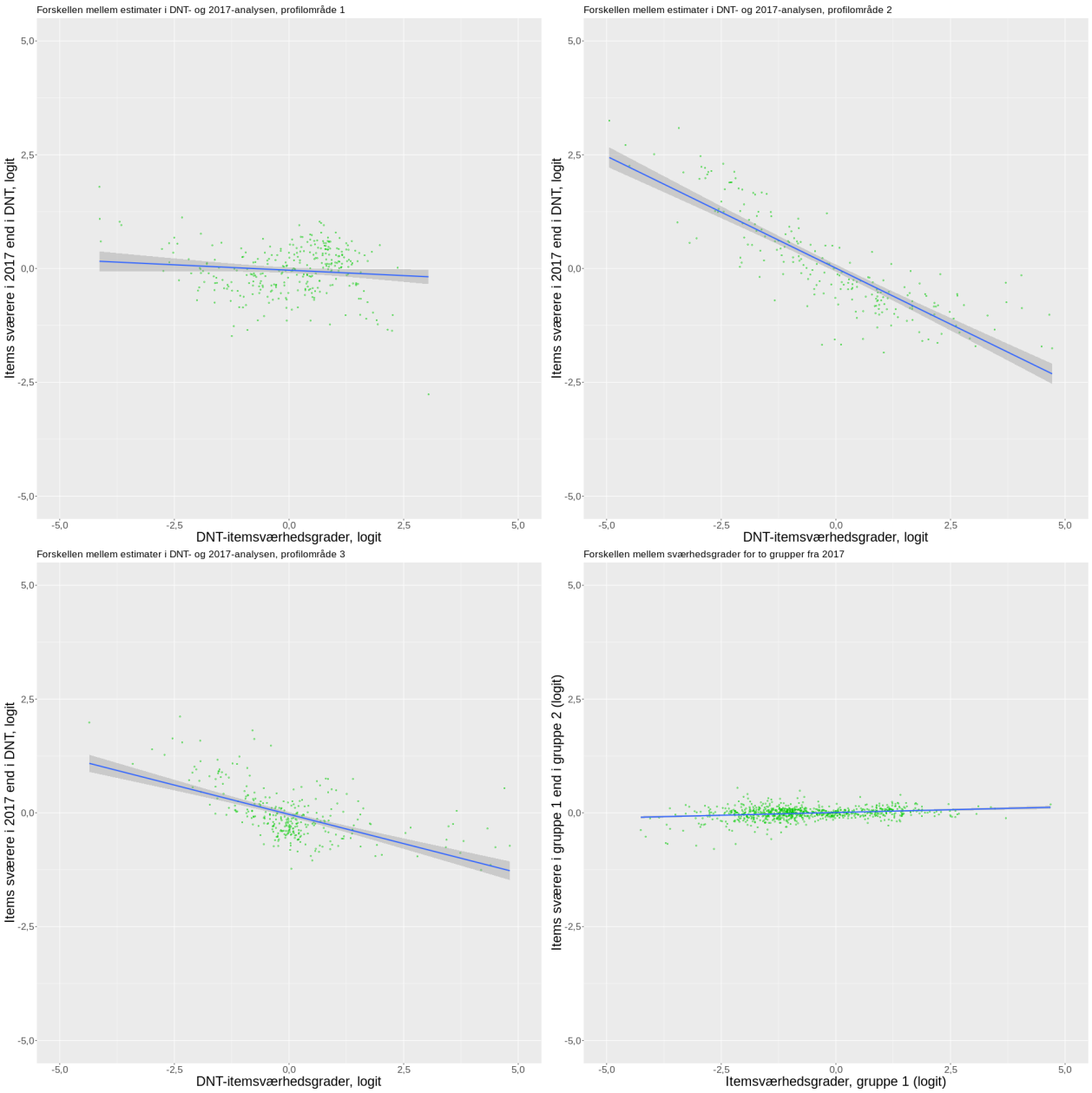
Figur 4.1 viser sammenhængen mellem sværhedsgraderne ifølge 2017-analysen og DNT. Det fremgår tydeligt af figuren at DNT- og 2017-analysen ikke er helt enige om vurderingen af sværhedsgraden af de enkelte items, men punkterne fordeler sig om den sorte identitetslinje for alle tre profilområder. Figurerne illustrer de høje korrelationer[3] mellem de to sæt sværhedsgrader hvilket betyder at DNT- og 2017-analysen er nogenlunde enige om hvilke opgaver der er mere eller mindre vanskelige.

For at give et indtryk af Rasch-modellens probabilistiske natur, og dermed af hvordan man må acceptere forskelle i estimater på tværs af analyser, har vi gennemført en analyse af data hvor vi har delt eleverne op i to halvdele med henholdsvis 24.240 og 24.241 tilfældigt valgte elever. For hver af disse grupper har vi kørt den samme Rasch-analyse som for hele gruppen af elever, og derefter har vi sammenlignet estimaterne af itemsværhedsgrader i de to analyser. Resultatet er gengivet grafisk (for alle tre profilområder i samme diagram) i det nederste højre diagram af figur 4.1.

[3] De er henholdsvis 0,91, 0,87 og 0,9 for de tre profilområder



Figur 4.1 Sammenligning af 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder. Der er indtegnet en sort identitetslinje og en blå linje med konfidensinterval der viser regression af 2017-værdier på DNT-værdier. Det nederste højre diagram viser estimater af sværhedsgrader (alle profilområder) for to tilfældigt valgte grupper elever fra 2017.

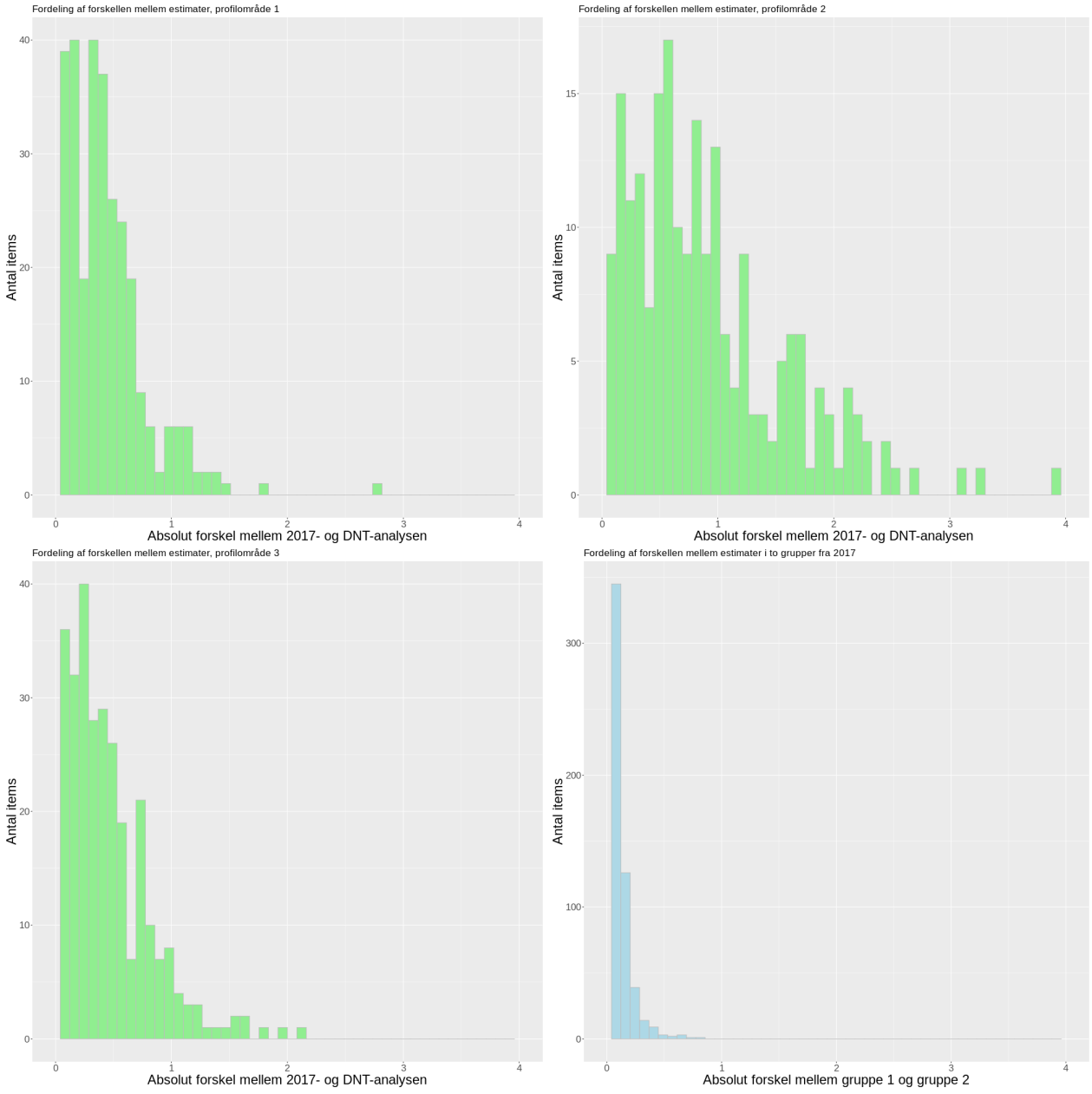


Figur 4.2 Forskellen mellem 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder i forhold til sværhedsgraden i DNT. Det nederste højre diagram viser forskellene i sværhedsgrader (alle profilområder) estimeret for to tilfældigt valgte grupper elever fra 2017.

I figur 4.2 er forskellen mellem DNT's og 2017-analysens estimat af itemsværhedsgrader plottet som funktioner af DNT-sværhedsgraderne. De faldende regressionslinjer viser at det for profilområde 2 og 3 gælder at de opgaver der var svære i DNT's estimation, generelt er blevet lettere, mens de opgaver der var lette, er blevet sværere. Sagt på en anden måde, er spredningen af opgavernes sværhedsgrader i profilområde 2 og 3 blevet mindre i 2017[4]. Særligt for profilområde 2 er forskellen dramatisk. Nederste højre diagram i figur 4.2 viser til sammenligning forskellene i analysen af to grupper elever fra 2017 som omtalt ovenfor.

[4]I første udgave af rapporten fremstod det som om disse forskelle også kendetegnede profilområde 1, men som det fremgår af figuren øverst til venstre, er det ikke tilfældet ifølge de korrekte itemsværhedsgrader

Histogrammerne i figur 4.3 viser fordelingen af den absolutte forskel på de to estimater af sværhedsgraden for de tre profilområder. Den gennemsnitlige absolutte forskel er henholdsvis 0,42, 0,91 og 0,44 for de tre profilområder hvilket på logitskalaen er udtryk for betydelige forskelle.



Figur 4.3 Fordelingen af forskellen mellem 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder, samt nederst til højre en sammenligning af estimat af sværhedsgrader (alle profilområder) målt på to grupper af elever fra 2017.

I tabel 4.2 ses antallet af items der ligger inden for intervaller af forskelle mellem DNT- og 2017-analysen.

I figur 4.4 ses de samme forskelle illustreret ved at angive de estimerede itemsværhedsgrader ifølge henholdsvis DNT- og 2017-analysen på hver sin linje, og forbinde det enkelte items estimat med en linje. Hvis estimatet

Tabel 4.2 Fordeling af absolut forskel i estimerede sværhedsgrader mellem DNT- og 2017-analyserne.

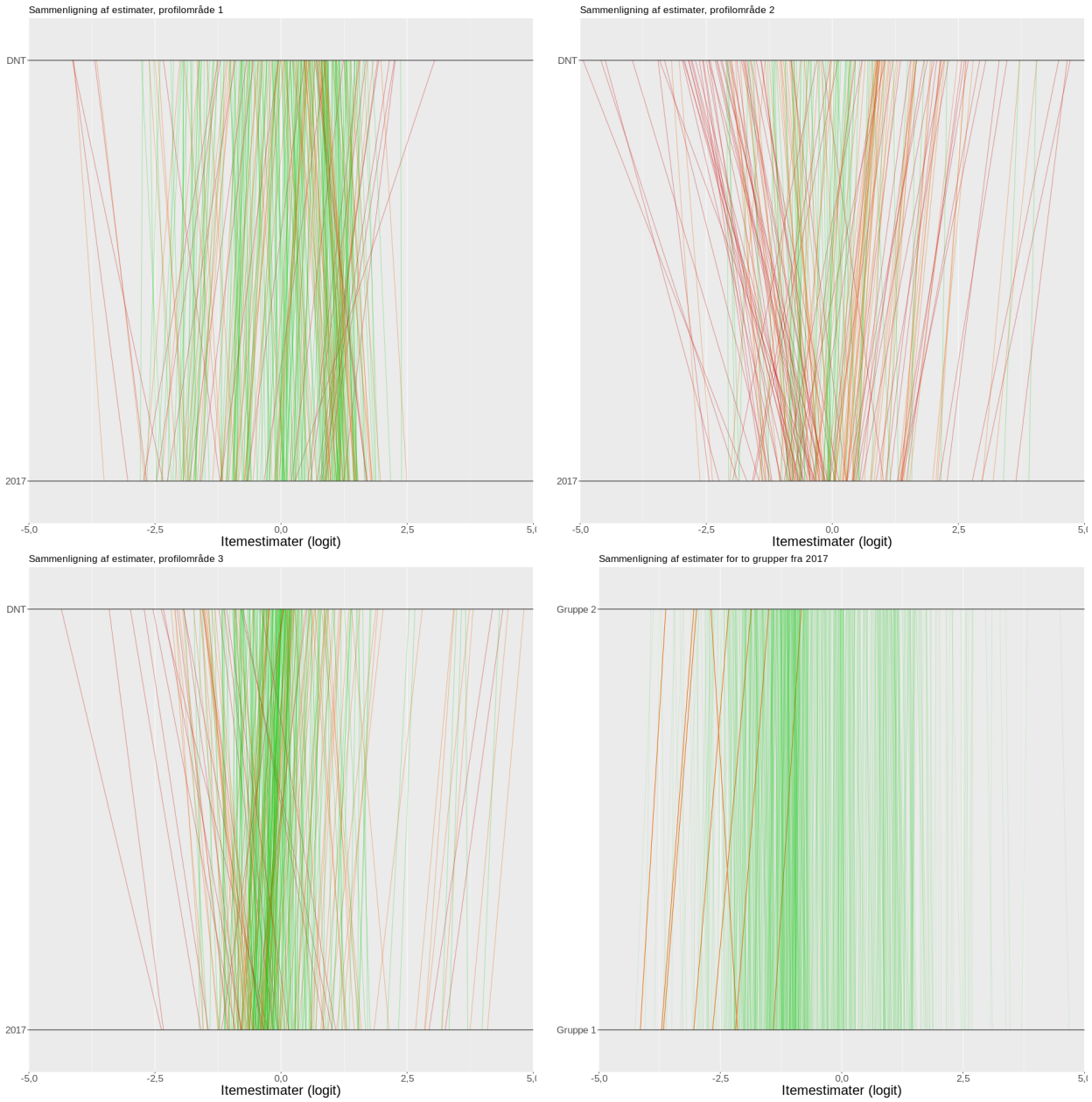
Interval	Profilområde 1	Profilområde 2	Profilområde 3
0 - 0,5	212	65	202
0,5 - 1	73	75	78
1 - 1,5	21	32	14
1,5 - 2	1	25	6
2 - 2,5	1	12	1
2,5 - 3		2	
3 - 3,5		2	
3,5 - 4		1	

Note:
Forskelle er medtaget i det interval hvis nedre grænse de er større end, og hvis øvre grænse de er mindre end eller lig med.

er nøjagtig det samme, vil linjen være lodret, og hvis der er forskel, vil linjen være skrå. Jo større hældning, des større er uenigheden om estimatet mellem de to analyser[5].

Det nederste højre diagram af figurerne 4.3 og 4.4 gengiver forskellene i itemsværhedsgrader i analysen af to grupper fra 2017. Som det fremgår, er der for nogle items uenighed om hvor svære items der overstiger 0,5 logit, er. Helt præcist drejer det sig om 8 items svarende til 1 procent. 4,4 procent af items estimeres mellem 0,25 og 0,5 forskelligt i de to analyser. Som det fremgår af disse tal såvel som af figurerne, er forskellene mellem 2017-analysen og DNT langt mere omfattende.

[5]Ideen til stregfigurerne er udviklet af Peter Allerup.



Figur 4.4 Sammenligning af 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder, samt nederst til højre en sammenligning af estimat af sværhedsgrader (alle profilområder) målt på to grupper af elever fra 2017. Der er grønne streger mellem estimater hvor forskellen er mindre end 0,5 logit, orange streger hvor forskellen er mellem 0,5 og 1 logit, og røde streger hvor forskellen er over 1 logit.

4.3.2 Betydning af at opgaver med forkerte sværhedsgrader vælges

Tabel 4.3 viser betydningen af at sværhedsgraderne i 2017 afviger fra de sværhedsgrader som nationale test-algoritmen anvender. Hvis nationale test-algoritmen vælger en opgave der ligger tæt på elevens dygtighed, vil chancerne for et korrekt svar være ca. 50 procent. Hvis opgavens sværhedsgrad målt på Rasch-modellens logitskala afviger med et point i 2017 i forhold til det nationale test-algoritmen har estimeret, vil chancerne for et korrekt svar være ca. 73 procent hvis opgaven er lettere, men kun 27 procent hvis opgaven er vanskeligere.

Tabel 4.3 Oversigt over betydning af forskellen på sværhedsgraderne i DNT og 2017 for opgaver, der ifølge DNT skulle svare til elevens dygtighed.

	Sandsynlighed for korrekt svar			
Absolut forskel	Lettere end antaget	Korrekt sværhedsgrad	Vanskeligere end antaget	Andel information ifht. optimalt
0,25	0,56	0,5	0,44	98 %
0,50	0,62	0,5	0,38	94 %
1,00	0,73	0,5	0,27	79 %
1,50	0,82	0,5	0,18	60 %
2,00	0,88	0,5	0,12	42 %
2,50	0,92	0,5	0,08	28 %
3,00	0,95	0,5	0,05	18 %
3,50	0,97	0,5	0,03	11 %

[1] Absolut forskel er forskellen mellem elevens dygtighed og opgavens sværhedsgrad uden fortegn.
[2] I søjlen 'Lettere end antaget' angives hvor stor sandsynligheden er for at eleven svarer rigtigt hvis opgaven er så meget lettere i 2017 som det er angivet i søjlen 'Absolut forskel'.
[3] I søjlen 'Korrekt sværhedsgrad' angives hvor stor sandsynligheden er for korrekt svar når opgaven har den forventede sværhedsgrad.
[4] I søjlen 'Sværere end antaget' angives hvor stor sandsynligheden er for et korrekt svar når opgaven er så meget sværere i 2017 som angivet i 'Absolut forskel'.
[5] I den sidste søjle angives hvor meget information der opnås når forskellen er så stor som angivet i 'Absolut forskel', sammenlignet med hvad den ville være hvis sværhedsgraden passede til elevens dygtighed.

Hvis forskellen er to point på logitskalaen, er der 88 procents chance for en lettere opgave og kun 12 procent for en vanskeligere opgave.

Hvis opgaven er lettere end DNT forventer, vil det naturligvis ikke være et problem for eleven, men det vil forringe sikkerheden af målingerne fordi et svar på en let opgave ikke bidrager med megen information der kan hjælpe til at præcisere vurderingen af målingen.

Det samme er tilfældet hvis opgaven er for vanskelig, men udover det vil eleven måske opleve testsituationen som ubehagelig fordi han skal løse en opgave der klart ligger ud over det han kan magte. Det forhold at den adaptive algoritme kan vælge opgaver der er vanskeligere end forventet og måske langt over det som eleven kan klare, kan måske forklare nogle af de dårlige oplevelser som lærere og forældre har rapporteret om i pressen, og som er beskrevet i undersøgelser (Bundsgaard og Puck 2016; Wandall, Nørrelund, og Nielsen 2018).

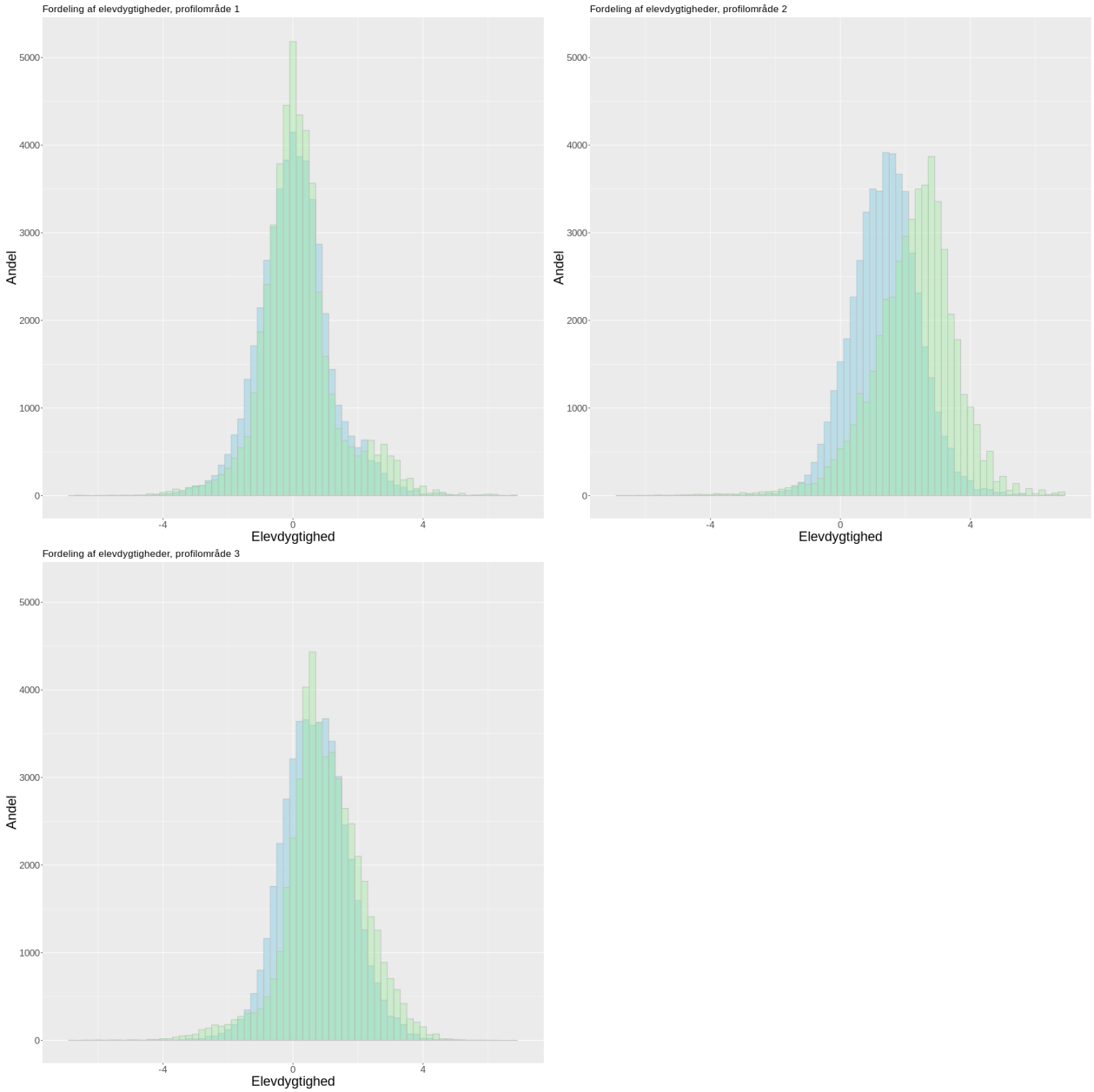
Den information som svaret på en opgave bidrager med, og som normalt omtales som item-information, kan kvantificeres. Det er derfor også muligt at måle præcis hvor megen item-information der mistes, hvis nationale test-algoritmen udvælger en opgave der afviger fra elevens dygtighed. Den sidste kolonne i tabel 4.3 giver en oversigt over disse forhold. Bemærk at forskelle på 0,5 kun fører til en beskeden reduktion af item-informationen. Denne iagttagelse har betydning for den adaptive algoritme som i sagens natur sjældent kan finde en opgave med en sværhedsgrad der passer præcist til elevens dygtighed. Hvis en udvalgt opgave rammer 0,25 logit ved siden af målt på logitskalaen, vil item-informationen kun være to procent mindre, end hvis algoritmen havde valgt en opgave med præcis den samme sværhedsgrad som elevens dygtighed.

4.4 Fordeling af elevdygtighed

De følgende figurer beskriver fordelingen af dygtigheden sådan som DNT ser den, og sådan som den i virkeligheden er ifølge 2017-analysen[6].

[6] Af denne figur fremgår det at antagelsen fra MCML-analysen om normalfordelte elevdygtigheder bekræftes af 2017-analysen og anvendelse af marginale estimater er derfor velbegrundet.

I figur 4.5 er fordelingen af elevdygtigheder ifølge 2017- og DNT-analyserne gengivet som histogrammer der er lagt ind over hinanden således at man kan se hvor de adskiller sig. DNT's elevdygtigheder er hentet fra datasættet som den estimerede dygtighed efter det sidste item, eleven har besvaret.



Figur 4.5 Fordelingen af elevernes dygtigheder. Den blå nuance angiver 2017-analysens estimat, og den grønne nuance angiver DNT's estimat.

Begge fordelinger har ved et umiddelbart blik de karakteristiske træk ved en normalfordeling i form af to haler for de dygtige og de mindst dygtige elever og en stor krop hvor flertallet af eleverne befinder sig.

Men det kan også se ud som om fordelingerne i DNT's estimerede elevdygtigheder er præget af en vis skævhed. Således er der tilsyneladende flere under end over middel dygtige elever i profilområde 2, mens det kan se ud som om der er flere over middel end under middel dygtige elever i profilområderne 1 og 3.

Tabel 4.4 Centrale parametre for beskrivelse af tilnærmelsen til normalfordelingen for 2017-analysens og DNT's estimerede elevdygtigheder.

	Middelværdi		Median		Standardafvigelse		Kurtosis		Skævhed	
Dimension	2017	DNT	2017	DNT	2017	DNT	2017	DNT	2017	DNT
Profilområde 1	0,05	0,18	0,03	0,07	1,1	1,2	1,46	2,6	0,17	0,53
Profilområde 2	1,38	2,26	1,40	2,39	1,1	1,3	1,53	3,2	-0,08	-0,73
Profilområde 3	0,70	0,97	0,68	0,92	1,0	1,2	0,82	1,9	-0,06	-0,42

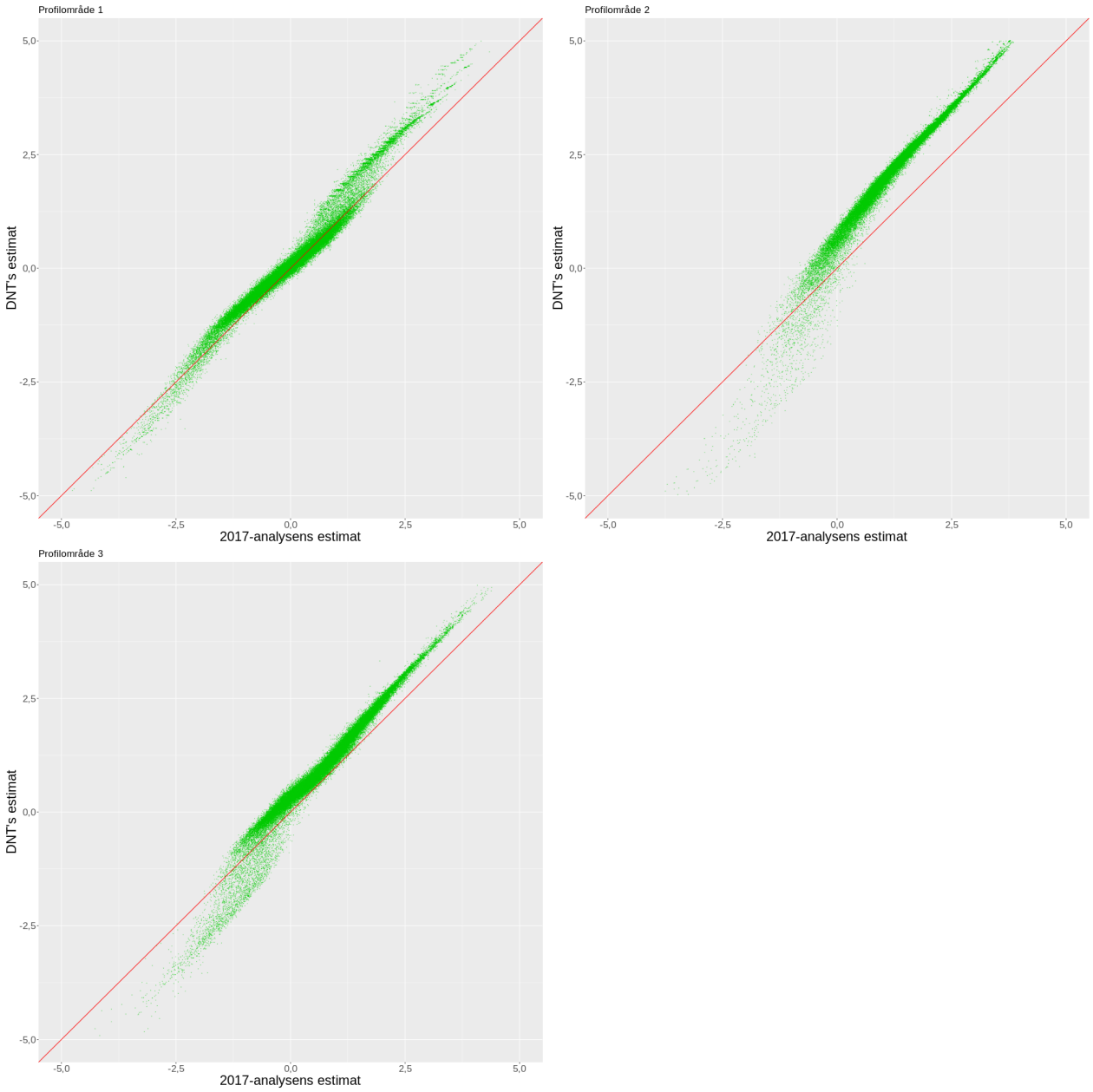
Dette underbygges af parametrene i tabel 4.4 som indeholder de centrale parametre til beskrivelse af de to fordelingers overensstemmelse med normalfordelingen.

Som det fremgår er Kurtosis for både 2017- og DNT-analysen højere end 0 som er værdien for en ren normalfordeling. Dette er udtryk for at der findes en del outliers i fordelingerne hvilket kan forventes idet nogle elever vil svare ukoncentreret, blive stoppet før tid, osv. Men forskellen på 2017- og DNT-fordelingen er væsentlig, og DNT identificerer således væsentligt flere outliers end 2017-analysen.

Forskellen i skævhed i 2017- og DNT-fordelingerne er endnu mere iøjnefaldende. For 2017-analysen er skævheden tæt på 0 for alle tre dimensioner, mens den for DNT er høj, særligt for profilområde 2. Dette underbygger de iagttagelser der kunne gøres ved blot at betragte histogrammerne.

4.5 Estimater på dygtigheder i 2017-analysen og i DNT's analyse

Figur 4.6 viser en sammenligning af estimatet af dygtigheden for de enkelte elever ifølge 2017-analysen og DNT.



Figur 4.6 Sammenligning af 2017-analysens og DNT's estimat af elevdygtigheder for de tre profilområder. Dygtighederne ifølge 2017-analysen findes på x-aksen og dygtighederne ifølge DNT's analyse findes på y-aksen. De grønner prikker symboliserer hver for sig én elev. Den røde streg er en identitetslinje der viser hvor målingerne er ens i de to analyser.

Som det fremgår er de to målinger af dygtigheder positivt korreleret, men de bekymrende tendenser som blev identificeret i undersøgelsen af fordelingen af elevdygtighederne (afsnit 4.4), bliver meget tydelige i disse figurer. Det gælder for alle tre profilområder at de mindst dygtige elever vurderes for lavt af DNT (grønne prikker ligger under identitetslinjen i den nederste del af dygtighedsskalaen), og de dygtigste vurderes for højt (grønne prikker over identitetslinjen). Særligt for profilområde 2 er det en meget markant tendens, men også profilområde 3 er den betydningsfuld. Nationale test producerer således systematisk forkerte estimater af elevernes dygtighed.

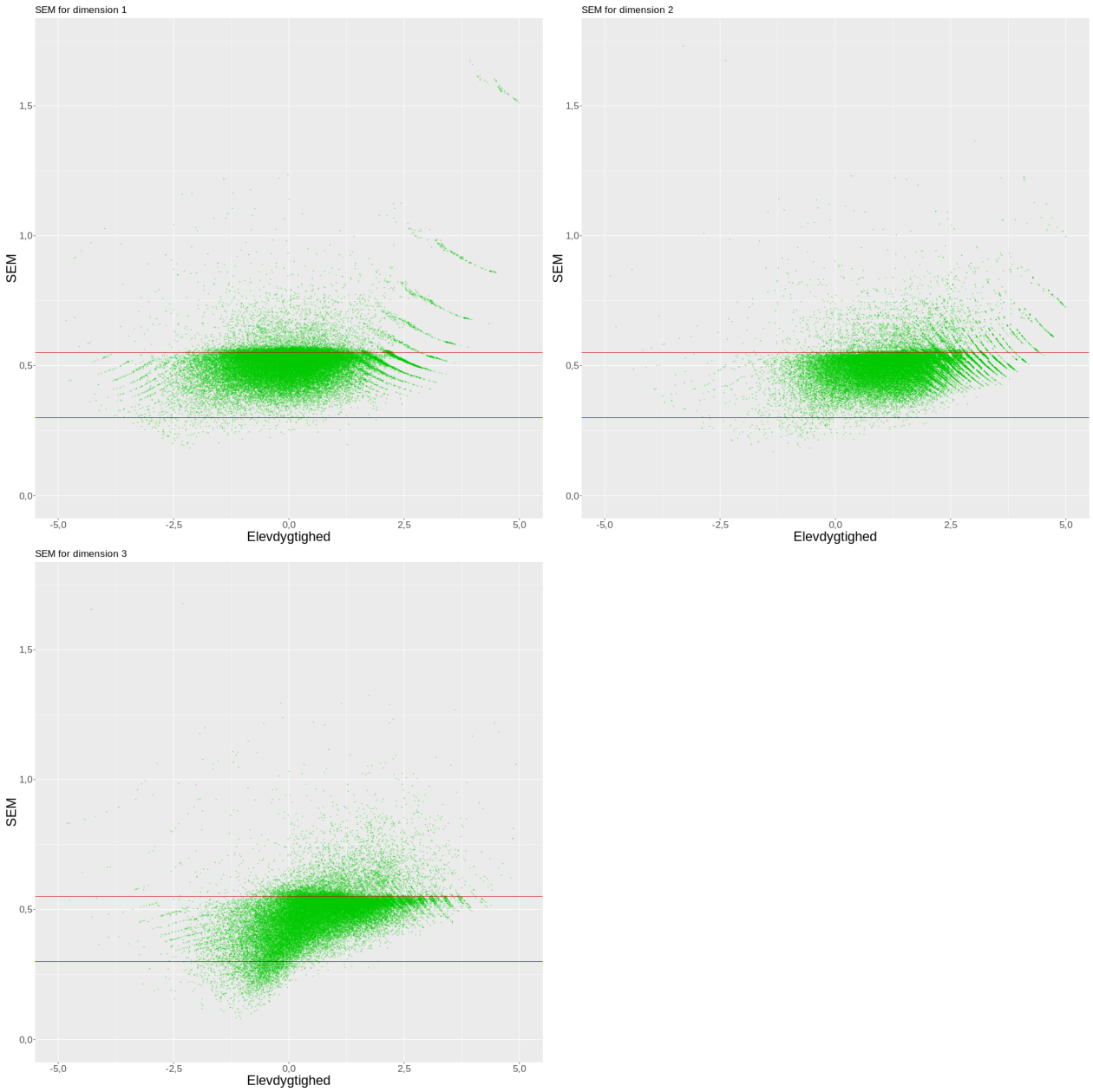
4.6 Standard Error of Measurement (SEM)

Standard Error of Measurement (SEM) er et udtryk for hvor sikker man kan være på den enkelte elevs estimerede resultat. Det der gør en adaptiv test eftertragtelsesværdig, er at den ideelt set stiller opgaver på elevens dygtighedsniveau og derved får meget information ud fra elevenes svar, således at alle elevers resultat måles med den samme sikkerhed. Som tidligere beskrevet (i afsnit 3.4) anvendes SEM af den adaptive algoritme som kriterium for hvornår estimatet på elevens dygtighed er tilstrækkeligt præcist, og testen derfor kan afsluttes.

I et af de dokumenter der lå til grund for udviklingen af nationale test, Svend Kreiners notat om adaptive test (Kreiner 2007), omtales en SEM på 0,16 som ønskværdig og realistisk i en perfekt adaptiv test. SEM afhænger af antallet af opgaver der stilles. I udviklingsprocessen blev det klart at det ikke var realistisk med et antal opgaver som kunne få SEM ned på 0,16. I stedet blev kravet til nationale tests målinger at de skulle have en SEM på 0,3 hvilket generelt betragtes som en god sikkerhed som vil resultere i en acceptabel reliabilitet på 0,9, hvis variansen af tallene for dygtigheden er tæt på 1.

Det lader til at det i de første tre-fire år af testens levetid var forventningen hos ministeriet at der faktisk var tale om en SEM på 0,3 (Wandall 2010). Men i 2014 blev det klart for ministeriet at algoritmen ikke var programmeret til at stoppe ved en SEM på 0,3, men allerede ved en SEM på 0,55 (Ravn 2014). Ifølge Undervisningsministeriets oplysninger betragter den adaptive algoritme således forløbet i et profilområde som afsluttet når SEM når under 0,55 (Styrelsen for It og Læring 2015). Derefter giver den adaptive algoritme kun opgaver i de profilområder der ikke er nået under 0,55. Når SEM er nået under 0,55 i alle tre profilområder, begynder algoritmen at give opgaver inden for alle tre områder igen, og læreren får med en grøn mærkat besked om at testen kan afsluttes.

I figuren 4.7 vises SEM som den er beregnet i 2017-analyserne, som funktion af elevdygtigheden (theta) i de tre profilområder.



Figur 4.7 SEM i forhold til elevdygtighed.

Der er lagt en linje ind ved 0,3 (blå) og 0,55 (rød) logit for at illustrere hvor mange elever der ligger over de grænseværdier der har været nævnt i debatten om nationale test. Som det fremgår, når en stor del af eleverne ikke at få en SEM under de forventede 0,55. Faktisk er det ganske store procentdele som ligger over dette skæringspunkt. For profilområde 1 har 16 procent af eleverne en SEM over 0,55 og 100 procent af eleverne en SEM over 0,3. Gennemsnittet er 0,51. For profilområde 2 har 12 procent af eleverne en SEM over 0,55 og 100 procent af eleverne en SEM over 0,3. Gennemsnittet er 0,5. For profilområde 3 har 13 procent af eleverne en SEM over 0,55 og 96 procent af eleverne en SEM over 0,3. Gennemsnittet er 0,48. I tabel 4.5 gengives fordelingen af SEM for de tre profilområder.

Til sammenligning har den amerikanske nationale test *Smarter Balanced* i English Language Arts/Literacy

Tabel 4.5 Fordeling af SEM for de tre profilområder (procentandele).

SEM-interval	Profilområde 1	Profilområde 2	Profilområde 3
0 - 0.3	0,4	0,4	4,3
0.3 - 0.4	5,2	5,4	12,8
0.4 - 0.5	35,0	38,0	33,5
0.5 - 0.55	43,3	44,0	36,8
0.55 - 0.6	11,3	7,1	7,4
0.6 - 0.7	2,9	3,4	3,5
0.7 - 1	1,6	1,5	1,6
>1	0,2	0,2	0,2

Note:
Intervallerne indeholder ikke den nedre værdi, men den øvre.

Tabel 4.6 Oversigt over sandsynligheder for korrekte svar for opgaver med en sværhedsgrad der svarer til dygtigheden.

	Sandsynligheder for korrekt svar	
SEM	I bunden af konfidensintervallet	I toppen af konfidensintervallet
0,20	0,40	0,60
0,30	0,36	0,64
0,40	0,31	0,69
0,50	0,27	0,73
0,55	0,25	0,75

Note:
Sandsynlighedernes beregnes som funktioner af SEM for yderpunkterne i 95 procent-konfidensintervallet omkring estimatet af dygtigheden.

for elever i 8. klasse en SEM der ligger mellem 0,30 og 0,41 med et gennemsnit på 0,32 (Smarter Balanced Assessment Consortium 2018, 2–58)[7].

En SEM på 0,55 er ganske betragtelig og langt ud over hvad man normalt ville acceptere for en pædagogisk test. Hvis estimatet af dygtigheden er tilnærmelsesvis normalfordelt, vil et 95 procent-konfidensinterval være defineret af dygtigheden $\pm 1,96 \times SEM$, og den sande dygtighed vil således med 95 procets sandsynlighed ligge inden for et interval hvor afstanden fra den mindste til den største værdi er lig med $3,92 \times SEM$. For at fortolke disse værdier kan man se på sandsynlighederne for korrekte svar på en opgave, der svarer til estimatet af dygtigheden, hvis den sande værdi var nede i bunden og oppe i toppen af konfidensintervallet. Tabel 4.6 viser disse sandsynligheder for forskellige værdier af SEM.

Udover at konkludere at en SEM-værdi på 0,20 fortæller nogenlunde samme historie om eleven i bunden og toppen af konfidensintervallet og helt forskellige historier hvis SEM er 0,55, vil vi overlade det til læseren at vurdere hvilke SEM værdier der skal betragtes som acceptable. Vi vil dog opfordre til at der gennemføres studier af hvilken størrelse af der kan accepteres under praktiske forhold. Sådanne studier kunne tage udgangspunkt i læreres begrundede vurdering af hvilke grupper af elever der adskiller sig dygtighedsmæssigt i en klasse, og sætte som mål at testen kan identificere et tilsvarende antal grupper som lærerne er i stand til.

[7]SEM er opgivet på den transformerede skala Smarter Balanced bruger (ibid. s. 5-4). Med en hældning på 85,8 svarer en SEM på 27,6 til 0,41 på logitskalaen.

4.6.1 Konfidensintervaller for percentilscorer

Nationale test resultatside til lærerne viser elevernes dygtighed på en såkaldt percentilskala (som relaterer sig til en historisk population, ikke den aktuelle som eleven er en del af). Denne skala er valgt for at give læreren et forståeligt mål for elevens dygtighed, men den indebærer at elever i det midterste interval af skalaen vil have forsvindende små forskelle i estimeret dygtighed på logitskalaen (jævnfør fordelingen af elever i figur 4.5). Det betyder at usikkerhedsintervallet vist på percentilskalaen vil forekomme utrolig stort.

En beregning af percentilscorer for den undersøgte population (ikke den historiske som anvendes til at angive percentilscorer i nationale test) findes i bilag D.

For at finde konfidensintervallet for en elevs percentilscore skal man lægge konfidensintervallet om estimatet af dygtigheden på logitskalaen og derefter beregne percentilværdierne for konfidensintervallets yderpunkter.

Hvis dygtigheden er lig med 0 på logitskalaen, og hvis SEM er 0,55, skal grænserne for konfidensintervallet gå fra percentilværdien svarende til ca. -1,0 på logitskalaen til percentilværdien for ca. +1,0. I 2017-analysen for sprogforståelse svarer det et konfidensinterval for percentilscoren fra ca. 16 til ca. 84[8].

Årsagen til det brede konfidensinterval er let at gennemskue når man ser på percentilkurverne til højre på figur 4.8. Alle kurver starter og slutter med at være meget flade, men er særdeles stejle i den største del af forløbet inde omkring logit-værdier på 0. Hvis man lægger konfidensintervallet på logitskalaen omkring værdierne -2,5 eller 2,5, vil konfidensintervallerne omkring percentilværdierne være meget snævrere.

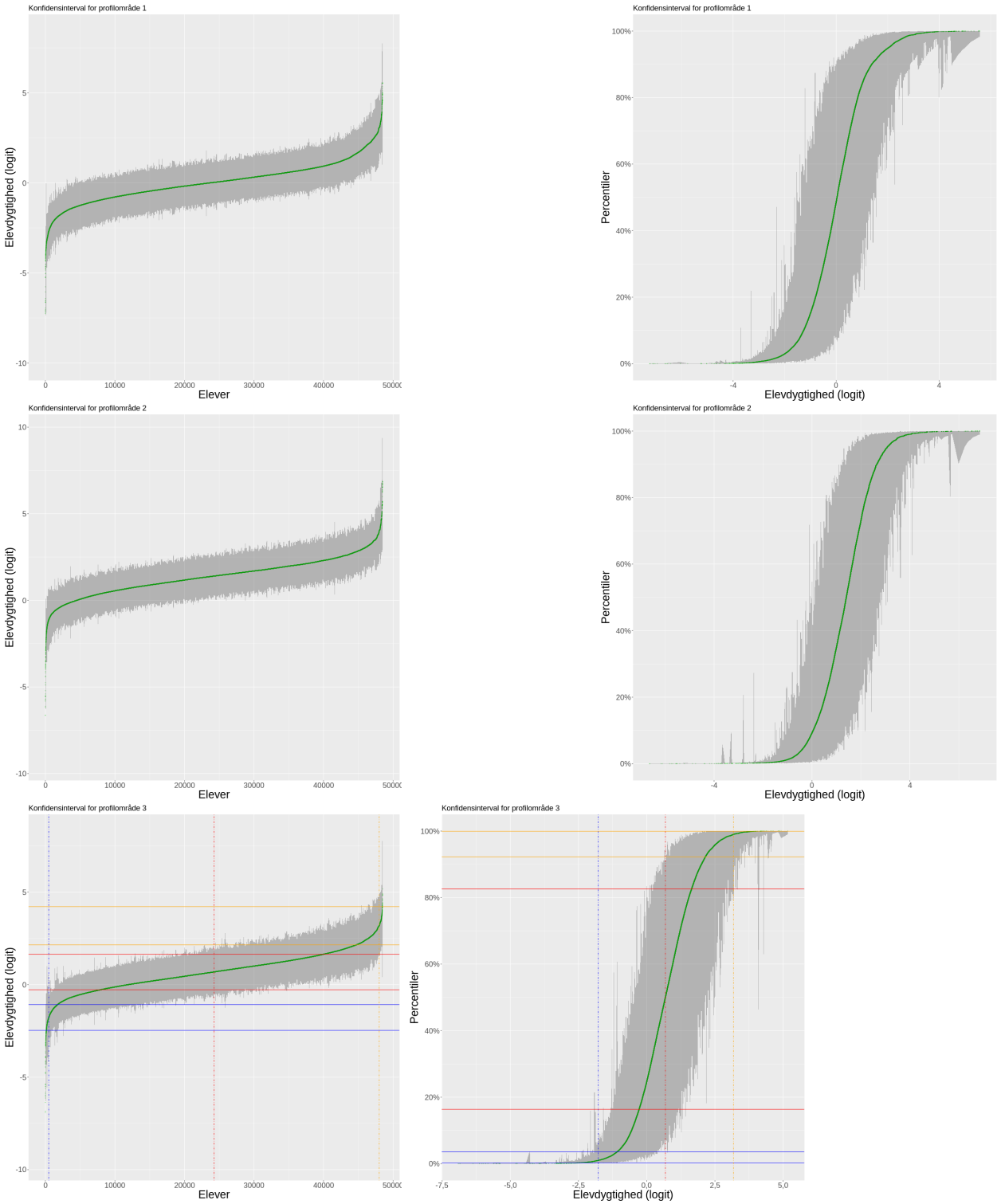
Figur 4.8 viser konfidensintervaller for logit- og percentilskalaen. Det fremgår af diagrammet til venstre at eleverne generelt har et konfidensinteraval af nogenlunde samme størrelse målt på logitskalen. Diagramet til højre viser at percentilskalaens konfidensinterval er anderledes stort inde omkring midten. Bæltet omkring kurven snævrer ind jo længere væk man komme fra midtpunktet, og er nærmest eksploderet inde omkring midten.

Denne eksplosion bliver tydelig i det tredje plot, hvor vi har fremhævet konfidensintervallet for tre elever henholdsvis øverst i den første percentil (konfidensintervallet er mellem de blå horisontale linjer), i den 50. (mellem de røde linjer) og den 99. percentil (mellem de orange linjer). SEM for de tre elever er henholdsvis `0,27`, `0,51`, `0,31`, hvilket for alle tre elever er under gennemsnittet af SEM. De stiplede streger angiver henholdsvis elevens placering i forhold til andre elever og elevens dygtighed på logitskalaen.

I det venstre plot ses det at usikkerhederne er betydelige på logitskalaen, men også at de er af samme størrelsesorden på logitskalaen. Men når man ser på percentilværdierne, som jo er et udtryk for eleverne ligger i forhold til andre elever, bliver det tydeligt at fordi der er så mange elever omkring midten af dygtighedsintervallet, så vil percentilscoren have et meget stort konfidensinterval knyttet til sig.

Ved beslutningen om at indføre nationale test valgte man at anvende en adaptiv test i stedet for en simpel lineær test fordi adaptive test måler færdigheder med samme sikkerhed for alle elever. Figur 4.8 viser konsekvensen af at man samtidig valgte at rapportere testresultaterne på en måde der tilsidesætter denne fordel og giver resultater der for langt de fleste elever er så usikre at man må fraråde lærerne at lægge vægt på dem. Percentilscorer kan kun have betydning for beslutninger om meget svage elever og meget stærke elever.

[8] I Styrelsen for It og Lærings notat *De nationale tests måleegenskaber* (Styrelsen for It og Læring 2016c) angives det at gennemsnittet af 95 procent-konfidensintervallet er ±12,5 percentilpoint. Og i en figur angives det at intervallet omkring 50. percentil er ca. ±17. På baggrund af vores analyser af nationale test for dansk, læsning ligger intervallet snarere omkring ±34.



Figur 4.8 Elevdygtigheder med 95 procent-konfidensinterval ifølge 2017-analysen. Venstre side viser konfidensintervallerne for hver elev på hver af de tre profilområder på logitskalaen ifølge 2017-analysen. For at bestemme en elevs score med konfidensinterval finder man eleven på x-aksen (eleverne er sorteret efter stigende dygtighed). Så går man op til den grønne kurve, og her er estimatet af elevens dygtighed på y-aksen. Det grå område rundt om kurven er 95 procent-konfidensintervallet. Højre side viser konfidensintervallet på percentilskalaen.

4.6.2 Kriteriebaserede scorer

De nationale test beregner elevernes dygtighed ved hjælp af estimater af Rasch-modellens personparameter på samme måde som det for eksempel er tilfældet i PISA og andre undersøgelser der anvender pædagogiske test. Disse parametre betragtes af mange som utilgængelige abstrakte størrelser. I stedet for at benytte disse værdier rapporterer nationale test resultaterne i form af percentilscorer med reference til fordelingen af elever i 2010 eller 2014 og i form af såkaldte kriteriebaserede scorer som har til hensigt at informere lærere og forældre om i hvilken grad fageksperter anser testresultaterne som mere eller mindre tilfredsstillende.

Percentilscorer og den usikkerhed der er knyttet til disse målinger, blev omtalt i forrige afsnit. Med hensyn til de kriteriebaserede scorer henvises til Bilag E hvor der er eksempler på hvorledes de rapporteres, og hvor man kan se at de kriteriebaserede scorer klassificerer testresultatet i en ud af følgende syv kategorier

1) Fremragende
2) Rigtig god
3) God
4) Jævn
5) Mangelfuld
6) Ikke tilstrækkelig

De betegnelser som nationale test knytter til de kriteriebaserede scorer, minder påfaldende om de korte betegnelser der er knyttet til det karaktersystem som anvendes i dag. I hvor høj grad dette er et tilfælde, er det ikke muligt at sige fordi ministeriet aldrig har dokumenteret præcis hvorledes kategorierne er defineret, og præcis hvad det er der fx karakteriserer et rigtig godt, men alligevel ikke fremragende præstation i nationale tests læseprøve.

Udgangspunktet for vurderingerne er at man har bedt fageksperterne om at udvælge et passende antal opgaver med varierende sværhedsgrader som eksperterne dels anså for gode opgaver, og som tilsammen dækker de færdighedsniveauer som elever på det pågældende klassetrin må forventes at ligge på. Wilson (2005) omtaler sådanne beskrivelser af opgaver og færdighedsniveauer som *construct maps*.

Det er som sagt uklart hvordan de seks niveauer er defineret, men hvis eksperterne har gjort det på den rigtige måde, har de gjort følgende:

a) Hver af de kriteriebaserede kategorier er defineret som et interval på skalaen med værdierne af Rasch-modellens personparametre. Den øvre grænse i et sådant interval angiver de dygtigste elever i kategorien, mens den nedre grænse referer til de mindst dygtige.
b) For både de dygtigste og de mindst dygtige beregnes først sandsynlighederne for at de udvalgte opgaver besvares korrekt ud fra personparametrene og sværhedsgraderne for opgaverne.
c) Ud fra disse sandsynligheder kan man dels beregne den samlede score på de udvalgte opgaver og dels identificere opgaver der er alt for vanskelige for eleverne i intervallet, og opgaver der er så lette at man kan argumentere at eleverne i det pågældende interval er kommet forbi og har styr på den type af problemer som opgaverne indeholder.

Fageksperternes kategoriseringer af præstationerne fra ”Fremragende” til ”Ikke tilstrækkelige” bør være og er formodentlig baseret på beregningerne i pkt. c. På grund af hemmelighedskræmmeriet omkring opgaverne er det imidlertid ikke muligt for fageksperterne at referere til konkrete opgaver fra testen hvilket i sagens natur gør det vanskeligt at dokumentere hvordan kategorierne er defineret, på en måde som lærere kunne have nytte af.

Uanset dette problem afslører analyserne i denne rapport at der kan være problemer med kategorierne fordi beregningerne af sandsynlighederne i pkt. b er baseret på forkerte oplysninger om opgavernes sværhedsgrader. Nogle opgaver, som man har troet var for vanskelige for elever på et bestemt niveau, er det måske ikke, mens andre opgaver, som så ud til at være meget lette for elever på et bestemt niveau, måske alligevel indeholder udfordringer for eleverne.

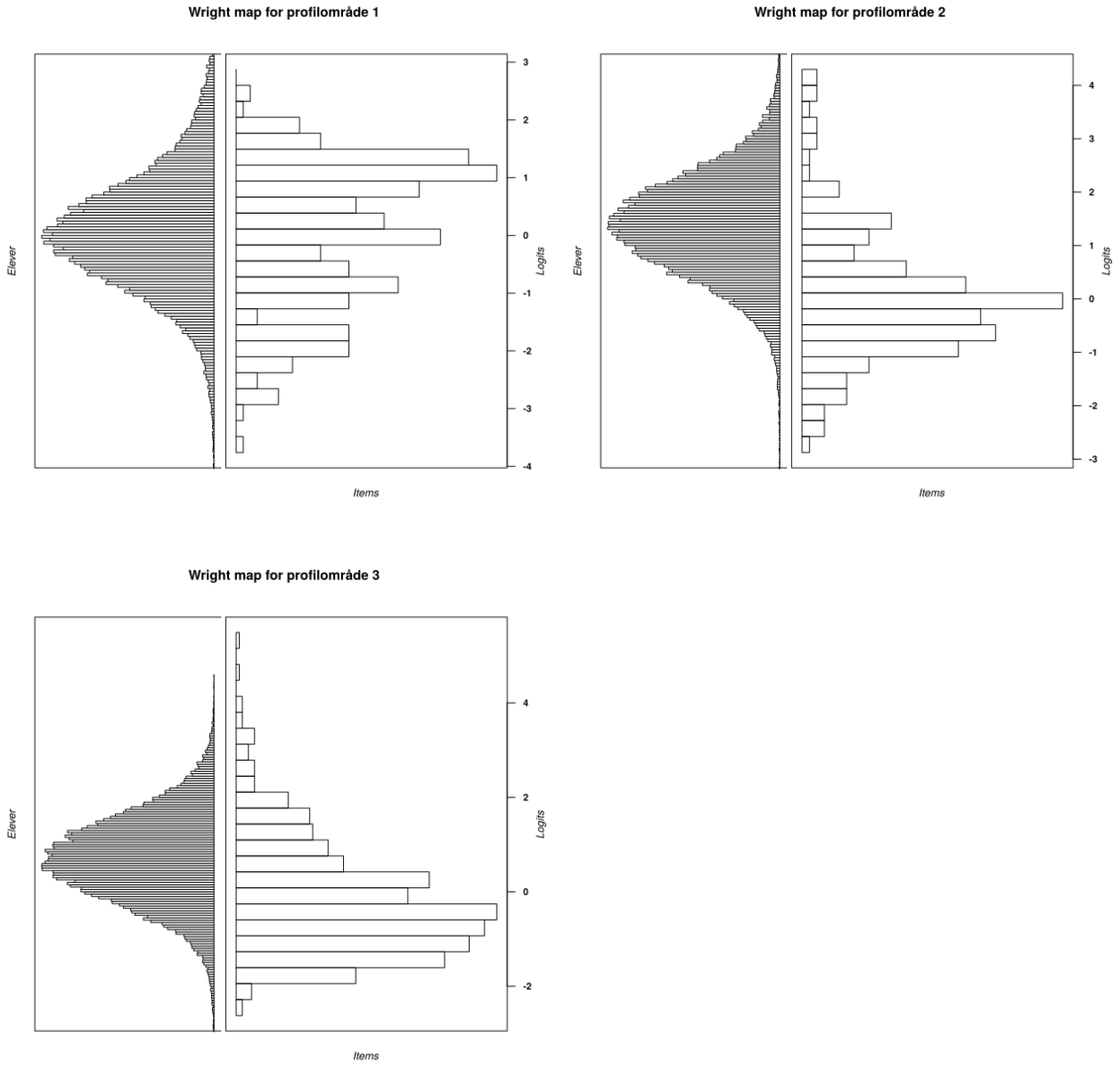
Uden kendskab til hvilke opgaver som fageksperterne har baseret deres vurderinger på, er det naturligvis ikke muligt for os at udtale os om der er store eller små problemer med de kriteriebaserede scorer som nationale test benytter. Men risikoen er der, og som konsekvens heraf må sandsynlighederne for de valgte opgaver genberegnes og kategorierne kontrolleres i forhold til de korrekte sandsynligheder.

4.7 Er der opgaver med passende sværhedsgrader til eleverne?

I almindelige lineært opbyggede tests hvor alle elever får de samme opgaver, vil man tilstræbe at have mange opgaver dér hvor man er interesseret i at måle elevernes dygtighed med stor præcision. Er det en diagnostisk test vil det betyde at man vil have mange opgaver omkring skæringspunktet for diagnosen. Er det en sammenlignende måling, vil man have mange opgaver i midten af fordelingen af elevernes dygtighedsniveau og færre ude i yderområderne.

Eftersom en adaptiv test tilstræber at måle alle elevers dygtigheder med stor præcision, fordrer det at der er et stort antal opgaver over hele dygtighedsspektret. Det er særligt antallet af lette opgaver og antallet af svære opgaver der er kritisk. Der skal være opgaver nok til både de allerdygtigste og de allerdårligste, men det gør ikke noget at mange elever får de samme opgaver. Til gengæld skal der, hvis man ønsker at teste igen, være opgaver nok til to eller flere runder uden at eleverne får den samme opgave to eller flere gange.

Et Wright-map (Wilson 2003) tegner fordelingen af sværhedsgraderne ved siden af fordelingen af elevernes dygtighed på den samme skala og viser derved om opgavernes sværhedsgrader dækker dygtigheden blandt eleverne. Se figur 4.9. Histogrammerne til venstre viser fordelingen af elevernes dygtigheder, mens histogrammerne til højre viser hvor mange opgaver der er på hvert sværhedsgradsniveau.



Figur 4.9 Wrightmaps for de tre profilområder.

Som det fremgår af de tre Wrightmaps, er tilpasningen mellem opgaver og elever ikke optimal for nationale test i dansk, læsning på 8. klassetrin. Profilområde 1, sprogforståelse, kommer tættest på idealet med et stort antal opgaver i midten og forholdsvis mange ude mod yderområderne, særligt i den høje ende. Havde testen været en almindelig lineær test, ville man betragte fordelingen af items på elever som yderst vellykket. Der er mange opgaver dér hvor der er mange elever.

Profilområde 2, afkodning, må betragtes som alvorligt underdimensioneret med opgaver i midten og den høje ende. Elever i midten af dygtighedsfordelingen og derover vil komme til at mangle opgaver på deres niveau, og derved vil de sandsynligvis opleve en stor mængde gengangere fra den frivillige til den obligatoriske test.

Også profilområde 3 har for få opgaver særligt hos de over middel dygtige elever.

Tabel 4.7 Korrelationer mellem dimensioner. Standardafvigelser i diagonalen.

	Profilområde 1	Profilområde 2	Profilområde 3
Profilområde 1	0,63	0,80	0,68
Profilområde 2	0,80	0,76	0,88
Profilområde 3	0,68	0,88	0,71

4.8 Et, to eller tre profilområder?

De foregående analyser giver anledning til at spørge om en af årsagerne til de problemer vi har observeret, skal findes i at testen ikke måler tre profilområder, men måske kun to eller et. Dette er et interessant og omdiskuteret teoretisk spørgsmål, men her vil vi blot undersøge det med statistiske metoder.

Til det formål er en række ekstra analyser af data foretaget. Analyser ved hjælp af flerdimensionelle Rasch-modeller (de såkaldte MCML-modeller) giver bedre muligheder for at undersøge sammenhænge på tværs af dimensioner (profilområder). Item-analysen baseres på en MCML-model der antager at værdierne på de forskellige dimensioner følger en fler-dimensional normalfordeling hvor værdierne på de valgte dimensioner antages at være korrelerede. I en multidimensionel analyse behandles således alle data i samme estimation, men det angives over for algoritmen at der er tale om et bestemt antal dimensioner således at der kan beregnes og tages højde for korrelationerne mellem dimensionerne. I tabel 4.7 ses en korrelationer mellem de tre dimensioner i nationale test.

Af tabellen fremgår det at der er ret stærke korrelationer mellem de tre dimensioner (profilområder). Når man er god til afkodning er man også god til sprogforståelse og tekstforståelse og omvendt. Det er af den grund naturligt at spørge om disse korrelationer skyldes at de tre profilområder måler forskellige aspekter af én og samme færdighed. Hvis det er tilfældet ville man kunne forenkle afrapporteringen af testresultaterne ved kun at beregne en samlet score som til gengæld – fordi den var baseret på svar på flere opgaver – ville være mere sikker (mindre SEM) end de tre individuelle profilresultater.

Analyser ved hjælp af MCML-modeller giver bedre muligheder for at afprøve hypoteser om endimensionalitet med flerdimensionelle modeller som alternativ, fordi modelstrukturen i sådanne hypoteser og alternativer er eksplicit defineret som en del af modellen i form af kovarianserne i den multivariate normalfordeling for profilområderne.

For at afklare om nationale test måler en, to eller tre forskellige dimensioner, er følgende hypoteser blevet afprøvet:

- Model 1: Alle tre profilområder måler én og samme færdighed.
- Model 2: Profil 1 & 2 måler samme færdighed. Profil 3 måler en særskilt færdighed.
- Model 3: Profil 1 & 3 måler samme færdighed. Profil 2 måler en særskilt færdighed.
- Model 4: Profil 2 & 3 måler samme færdighed. Profil 1 måler en særskilt færdighed.
- Model 5: De tre profilområder måler forskellige færdigheder.

For samtlige modeller skal itemsværhedsgrader og parametre for fordelingen af de bagvedliggende færdigheder estimeres for tilsammen 823 opgaver (1.156 items). Tabel 4.8 viser antallet af ukendte parametre for hver af de fem modeller samt de såkaldte *deviance*-værdier der skal bruges til at beregne *likelihood ratio*-test for en model som hypotese og en anden model som alternativ. Tabellen viser også to informationskriterier, der skal bruges til at vælge den bedste model, hvis resultaterne i den efterfølgende tabel skulle vise sig at være inkonsistente.

Da modellerne 1-4 alle er indlejret i Model 5, og da Model 1 er indlejret i modellerne 2-4, kan vi teste hypoteserne i tabel 4.9.

Tabel 4.8 Resultaterne af analyserne af fem modeller.

Model	Parametre	Deviance	AIC	BIC
Model 1: (Profil 1+2+3)	1157	3614960	3617274	3627443
Model 2: (Profil 1+2, Profil 3)	1159	3605957	3608275	3618461
Model 3: (Profil 1+3, Profil 2)	1159	3609154	3611472	3621658
Model 4: (Profil 2+3, Profil 1)	1159	3602171	3604489	3614676
Model 5: (Profil 1, Profil 2, Profil 3)	1162	3614331	3616655	3626868

[1] Deviance: Et mål for de empiriske datas afvigelse fra den teoretiske model.
[2] AIC: Akaike Information Criterion. Et mål for den relative kvalitet af en model.
[3] BIC: Bayesian Information Criterion. Et mål for den relative kvalitet af en model.

Tabel 4.9 Sammenligning af de fem modeller.

Hypotese	Alternativ	Likelihood-ratio	Frihedsgrader	p
Model 1	Model 2	9003	2	0
Model 1	Model 3	5806	2	0
Model 1	Model 4	12789	2	0
Model 1	Model 5	629	5	0
Model 2	Model 5	-8374	3	0
Model 3	Model 5	-5177	3	0
Model 4	Model 5	-12160	3	0

[1] Likelihood-ratio: Forholdet mellem plausibiliteten af to modeller.
[2] p: Sandsynlighed for forskel.

Det fremgår af tabellen at der er højsignifikant (p = 0) forskel på alle de sammenlignede modeller, og at hverken den unidimensionelle eller den tredimensionelle model beskriver data bedst (men den tredimensionelle er bedre end den unidimensionelle). Man kan ikke direkte sammenligne de tre alternative modeller via likelihood-ratioet, men AIC og BIC-værdierne er lavest for model 4, og dette kan bruges som argument for at profilområderne afkodning og tekstforståelse slås sammen. Dette ville dog være i modstrid med den tilgrundliggende teori om læsning: det såkaldte *Simple view of reading* (Hoover og Gough 1990)), der hævder at læsning (tekstforståelse) er lig med produktet af afkodning og sprogforståelse (*Læsning = Afkodning x Sprogforståelse*). Uoverenstemmelsen kan enten skyldes at de tre dimensioner ikke operationaliserer teoriens tre faktorer korrekt, eller det kan være et tegn på at teorien falsificeres med disse data. Vi går ikke yderligere ind i denne interessante iagttagelse i denne sammenhæng, men vil opfordre til at det undersøges nærmere.

4.9 Konklusioner og diskussion

Vi har i dette kapitel undersøgt data fra nationale test, dansk, læsning, i 2017 og sammenlignet resultaterne med nationale tests udsagn om itemsværhedsgrader og elevdygtigheder.

Vi har identificeret en række problemer:

1) De sværhedsgrader som anvendes i nationale test, er ikke længere korrekte i 2017.
2) Denne uoverensstemmelse fører til en uoverensstemmelse i estimering af elevdygtigheder, som kan

identificeres ved at se på fordelingen af elevernes dygtigheder som de er estimeret ud fra 2017-data, sammenlignet med dygtighederne estimeret af nationale test.

3) Uoverensstemmelsen viser sig at være systematisk således at nationale test undervurderer elever i den dårlige ende af skalaerne og overvurderer elever i den bedre ende af skalaerne, særligt i profilområde 2 og 3. Forskellen på dygtige og mindre dygtige elever er altså mindre end nationale test konkluderer.
4) Der er ikke tilstrækkeligt med opgaver for de dygtige elever i to profilområder, og i et profilområde mangler tillige opgaver til de middeldygtige elever.
5) Usikkerheden på estimatet af elevernes dygtighed er for stor, og for en ganske stor andel (ca. 15 procent) af eleverne er der tale om endnu større usikkerheder end den i forvejen høje grænse for SEM på 0,55. Problemet skyldes i et vist omfang at nationale test regner med forkerte sværhedsgrader, men problemet ville også være der hvis nationale tests sværhedsgrader var korrekte. SEM-værdier over 0,5 er i vores øjne uacceptable i forbindelse med individuel vurdering af testresultater.
6) Der er ikke belæg i data for at sige at de tre profilområder måler én og samme færdighed, men analysen tyder på at det kan give mening at se afkodning og tekstforståelse som én dimension.

De observerede forskelle på hvad DNT anvender som sværhedsgrader, og de sværhedsgrader som items faktisk har (2017-analysen), vil have betydelige konsekvenser for hvordan den adaptive algoritme opfører sig. For det første vil den adaptive algoritme vælge opgaver som reelt ikke passer til elevens dygtighed, og eleven vil derfor ikke svare som forventet. Det betyder at eleven vil være længere tid om at nå et tilstrækkeligt sikkert resultat. For det andet vil estimatet af usikkerheden hvile på forkerte itemsværhedsgrader, og algoritmen kan derfor stoppe for tidligt (eller for sent). Og endelig vil estimatet af elevens dygtighed hvile på forkerte itemsværhedsgrader og derfor være forkert. Det sidste er naturligvis det mest bekymrende.

Forskellen på fordelingerne baseret på DNT's beregninger og på 2017-analysen betyder at DNT tegner et forkert billede af fordelingerne af færdighederne i 2017 og især et forkert billede af fordelingen af de dygtigste og de mindst dygtige. Og det betyder at fordelingen af dygtigheden i 2014 sådan som DNT beregnede det, ikke kan sammenlignes med fordelingen i 2017. Konsekvensen af dette er at effekten af skolereformen ikke kan evalueres ved sammenligning af DNT's tal fra 2014 og fra 2017, således som det er forudsat i den politiske aftale fra 2013 ("Aftale Mellem Regeringen (Socialdemokraterne, Radikale Venstre Og Socialistisk Folkeparti), Venstre Og Dansk Folkeparti Om et Fagligt Løft Af Folkeskolen" 2013).

5

Analyser af enkeltstående testforløb

5.1 Indledning

Dette kapitel beskriver analyser af enkeltstående testforløb. Testforløbene er ikke udvalgt tilfældigt, men således at de kan illustrere to forskellige forhold.

For det første hvorledes den adaptive algoritme fungerer hvis man tager hensyn til opgavernes sande sværhedsgrader således som de er estimeret i 2017-analysen. I og med at analysen i kapitel 4 viste at der er meget store forskelle på de sværhedsgrader som den adaptive algoritme benytter, og de sande sværhedsgrader i 2017, men at der også var en høj grad af enighed om hvad der var let og hvad der var vanskeligt, forventes det at den adaptive algoritmes valg af nye opgaver stadig giver en vanskeligere opgave hvis den foregående opgave blev besvaret korrekt, og lettere hvis den blev besvaret forkert. Det forventes også at sværhedsgraderne af valget af opgaver ligger længere væk fra det aktuelle bud på dygtigheden, og at sikkerheden er dårligere end den usikkerhed der ville have været opnået hvis nationale tests oplysninger om sværhedsgraderne havde været korrekte.

For det andet vil det blive undersøgt om svarene i et enkelt testforløb fordeler sig på en måde der passer til en Rasch-model med sværhedsgraderne fra 2017. Årsagen til at dette problem tages op, er de mange fortællinger om "mærkelige" testforløb hvor lærere giver udtryk for en sund skepsis med hensyn til om målingerne af færdighederne stemmer overens med elevernes faktiske dygtigheder.

Fortællingerne om mærkelige testforløb bærer i mange tilfælde præg af at brugerne ikke helt kan gennemskue at testresultaterne er probabilistiske, og at mærkelige testforløb både kan og vil forekomme, men det betyder naturligvis ikke at mistroen er ubegrundet. For at undersøge dette problem har vi derfor udvalgt en række testforløb med fordelinger af svarene der svarer til de fortællinger man hører om, og vi vil beregne og vurdere sandsynligheder for at sådanne forløb kan forekomme hvis hele elevens testadfærd rent faktisk afspejler dygtigheden på den måde som Rasch-modellen og den adaptive algoritme forventer.

Mulighederne og nytteværdien af kontrol af enkeltstående testforløb og behovet for at der blev givet en advarsel, hvis der var et eller andet påfaldende ved forløbet, blev allerede diskuteret i forbindelse med udviklingen af nationale test uden at blive implementeret.

Det næste afsnit giver en kort beskrivelse af metoderne, hvorefter de følgende afsnit giver eksempler på anvendelserne og diskuterer konsekvenserne af resultaterne.

5.2 Metoder til at teste om enkelte testforløb kan beskrives ved hjælp af Rasch-modellen

Metoderne til afprøvning af om enkeltstående testforløb passer til en Rasch-model svarer til Fischers såkaldte eksakte test for uafhængighed i 2×2-tabeller.

I Fischers test beregnes den betingede sandsynlighed for tallene i tabellen givet række- og søjlesummerne under forudsætning af at de to variable er indbyrdes uafhængige. Denne sandsynlighed benyttes som

teststørrelsen ud fra den betragtning at en lille sandsynlighed er udtryk for at man har observeret noget uventet evt. grænsende til det usandsynlige hvis hypotesen om uafhængighed er korrekt. Da der kan være mange forskellige tabeller der giver de samme række- og søjlesummer, og da summen af sandsynlighederne for alle disse tabeller er lig med 1, følger det at den beregnede sandsynlighed kan være ganske beskeden uden at det i sig selv er udtryk for signifikant evidens mod hypotesen om uafhængighed. For at løse dette problem foreslog Fisher at beregne sandsynlighederne for samtlige tabeller med de givne række- og søjlesummer og at definere graden af signifikans som sandsynligheden for at observere en tabel der er lige så usandsynlig som eller mere usandsynlig end sandsynligheden for den observerede tabel.

Det er den samme måde vi vil benytte til at vurdere om det observerede svarmønster svarer til den adaptive algoritmes forventning om svar der passer til Rasch-modellen.

For hvert forløb beregnes den *betingede* sandsynlighed for svarmønstret *givet* sværhedsgraderne på items og det samlede antal point som eleven har opnået. Beregningerne af sandsynligheder kan være lidt komplicerede og vil ikke blive gennemgået her, men interesserede læsere kan finde flere detaljer i bilag B.

Vurderingen af om det observerede svarmønster er udtryk for signifikant evidens mod påstanden om at svarene kommer fra en Rasch-model, baseres på følgende sandsynligheder:

1) Sandsynligheden for at svarmønstret på opgaver fra en Rasch-model med de kendte sværhedsgrader giver et endnu mere usandsynligt svarmønster end det der rent faktisk er observeret. Selvom det observerede svarmønster i sig selv har en lille sandsynlighed for at forekomme, kan det jo godt tænkes at der er mange andre svarmønstre der er endnu mere usandsynlige, og hvis disse tilsammen har en stor sandsynlighed for at forekomme, kan man ikke påstå at det observerede svarmønster i sig selv er udtryk for manglende tilpasning til Rasch-modellen.
2) For hver enkelt opgave beregnes sandsynligheden for det givne svar givet det samlede antal korrekte svar på opgaverne. Hvis en elev har svaret korrekt på næsten alle opgaver, inklusive mange vanskelige opgaver, er et forkert svar på en meget let opgave udtryk for at der må være et eller andet der har forstyrret eleven mens denne opgave blev løst, hvis sandsynligheden for det givne svar er meget lille.
3) For hver eneste værdi af k beregnes sandsynligheden for at værdien af R_k er mindre end eller lig den observerede værdi, hvor R_k er lig med antal rigtige svar på de k første opgaver. Hvis denne sandsynlighed er lille, er det et udtryk for at man har observeret en usandsynlig dårlig start på forløbet af testen. I tabel 5.4 ses et forløb hvor $R_7 = 0$, og hvor sandsynligheden for at de første syv opgaver besvares forkert, når der tilsammen scores 13 ud af 21 mulige point, er mindre end 0,005. Et resultat der ville være usandsynligt hvis elevens dygtighed var kommet til udtryk på den måde som det er krævet af Rasch-modellen.

De to første sandsynligheder bruges til at vurdere signifikansen af henholdsvis det samlede svarmønster og svarene på de enkelte opgaver, men den tredje sandsynlighed bruges til at vurdere signifikansen af forløb med mange forkerte svar i starten eller slutningen af forløbet.

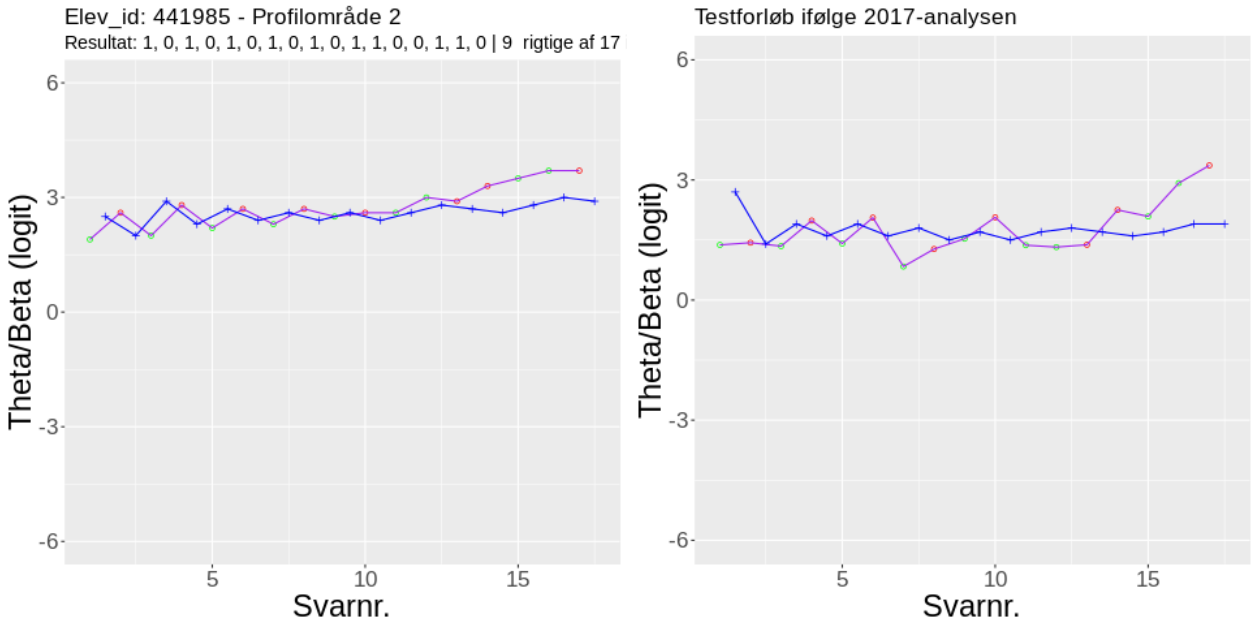
Eksemplet i de næste to afsnit vil illustrere metoderne for to forskellige forløb, hvorefter resultaterne for de øvrige forløb vil blive opsummeret og diskuteret.

5.3 Et adaptivt forløb

Først vises et eksempel på et forløb hvor eleven har svaret skiftevis rigtigt og forkert. Dette kan siges at være et tæt på ideelt adaptivt forløb som dog slutter u-adaptivt.

Tabel 5.1 Forløb for elev 441985, profilområde 2

Opgavenummer	Score	Kummuleret score	2017-sværhedsgrader	DNT-sværhedsgrader
0108020115072	1	1	1,38	1,9
010802000301234810-1	0	1	1,43	2,6
010802000301234966-1	1	2	1,35	2,0
010802000301234959-1	0	2	1,98	2,8
010802000301234953-1	1	3	1,41	2,2
010802000301234951-1	0	3	2,06	2,7
0108020111026	1	4	0,84	2,3
0108020110139-1	0	4	1,27	2,7
0108020111022	1	5	1,53	2,5
010802000301234848-1	0	5	2,07	2,6
0108020111019	1	6	1,37	2,6
010802000301238021-1	1	7	1,32	3,0
0108020111006	0	7	1,38	2,9
0108020110268	0	7	2,25	3,3
0108020110166-1	1	8	2,09	3,5
010802000301239338-1	1	9	2,92	3,7
010802000301239337-1	0	9	3,36	3,7



Figur 5.1 Testforløb for elev 441985, profilområde 2. Diagrammet til venstre viser data fra DNT's analyser af elevens forløb med anvendelse af DNT's itemsværhedsgrader. Rækkefølgen af de items eleven har besvaret, ligger på x-aksen. På y-aksen ses opgavernes sværhedsgrad. Hvis opgaven besvares korrekt, er den grøn, hvis den besvares forkert, er den rød. Hvis det er et polytomt item hvor nogle svar er korrekte, er prikken orange. Items er forbundet med en lilla linje. Den blå linje med blå krydser viser hvordan elevens dygtighed er estimeret efter den seneste opgave (derfor er punktet placeret mellem to items). Diagrammet til højre viser det samme baseret på 2017-analysens estimater.

Tabel 5.2 Oplysninger om testforløbet for elev 441985 baseret på 2017-sværhedsgraderne.

		Betingede sandsynligheder				Dygtighedsestimater		
itemno	opgavenummer	score	scoreprob	cumulated	cumulatedprob	WML	SEM	Next
1	108020115072	1	0,630	1	0,630			
2	010802000301234810-1	0	0,391	1	0,503			
3	010802000301234966-1	1	0,633	2	0,481	1,92	1,05	1,99
4	010802000301234959-1	0	0,528	2	0,405	1,55	0,99	1,42
5	010802000301234953-1	1	0,621	3	0,416	1,87	0,91	2,05
6	010802000301234951-1	0	0,546	3	0,365	1,61	0,83	0,87
7	108020111026	1	0,746	4	0,387	1,77	0,78	1,29
8	0108020110139-1	0	0,347	4	0,281	1,48	0,73	1,53
9	108020111022	1	0,593	5	0,358	1,69	0,69	2,09
10	010802000301234848-1	0	0,554	5	0,275	1,55	0,65	1,37
11	108020111019	1	0,634	6	0,359	1,70	0,62	1,34
12	010802000301238021-1	1	0,641	7	0,426	1,83	0,60	1,39
13	108020111006	0	0,370	7	0,297	1,65	0,57	2,28
14	108020110268	0	0,604	7	0,157	1,56	0,55	2,12
15	0108020110166-1	1	0,437	8	0,339	1,73	0,53	2,98
16	010802000301239338-1	1	0,238	9	0,831	1,93	0,52	3,39
17	010802000301239337-1	0	0,831	9	1,000	1,88	0,51	

Note:

[1] Scoreprob: Sandsynligheden for at eleven med den estimerede dygtighed vil svare som det skete, på denne opgave.
[2] Cumulatedprob: Sandsynligheden for at eleven med den estimerede dygtighed vil have haft det givne forløb indtil dette item.
[3] WML: Weighted Maximum Likelihood-estimat af elevens dygtighed på dette tidspunkt.
[4] SEM: Standard Error of Measurement for det estimerede resultat på dette tidspunkt
[5] Next: Næste opgaves sværhedsgrad.

Elev 441985 har, som vist i tabel 5.1, opnået ni point på 17 opgaver fra profilområde 2. Tabellen viser både sværhedsgraderne estimeret i datamaterialet fra 2017 og de sværhedsgrader som den adaptive algoritme benytter til estimation af elevens dygtighed. Bemærk at man ikke skal lægge vægt på den systematiske forskel der fortæller at DNT's sværhedsgrader er større end sværhedsgraderne i 2017. Det man skal lægge mærke til, er forskellen på sværhedsgraderne. Ifølge DNT's oplysninger er sværhedsgraden næsten den samme for de to sidste opgaver. Ifølge 2017-analysen er den sidste opgave den vanskeligste med en sværhedsgrad der ligger 0,40 over sværhedsgraden for den næstsidste.

Forløbet for elev 441985 er også vist i figur 5.1. Som det fremgår af den venstre side af figuren har algoritmen skiftevis estimeret elevens dygtighed højere og lavere og derved stillet en opgave der var sværere og lettere, og eleven har svaret skiftevis forkert og rigtigt (evt. to i træk). I et optimalt adaptivt forløb ville kurven med opgaverne ligge parallelforskudt en smule til højre for kurven med dygtigheden. Det sker i starten af forløbet, men mod slutningen begynder algoritmen at give opgaver der er væsentligt sværere end elevens dygtighed er vurderet til. Det kunne tyde på at der ikke er flere opgaver lige omkring elevens estimerede dygtighed. På figuren til højre er der endnu større uoverensstemmelse mellem estimeret dygtighed og efterfølgende opgave (omkring opgave 7, 8 og 12). Her bliver det tydeligt at de aktuelle sværhedsgrader på de items der tildeles, afviger fra estimatet af dygtigheden.

I tabel 5.7 sidst i dette kapitel er elevens dygtighed ifølge henholdsvis 2017-analysens og DNT's estimat gengivet som en percentilplacering.

I tabel 5.2 findes det samme testforløb suppleret med oplysninger der kan bruges til at vurdere om denne elev har besvaret opgaverne på en måde der svarer til Raschmodellens påstand om at resultatet kun afhænger af

elevens dygtighed og opgavens sværhedsgrad. For hver opgave er beskrevet den betingede sandsynlighed for scoren givet sværhedsgraderne (fra 2017) og elevens endelige scoreresultat, sandsynligheden for den kummulerede score (eller lavere) på det givne trin i forløbet, samt dygtigheden og SEM beregnet på det givne trin i forløbet. Den sidste søjle angiver sværhedsgraden af næste opgave i forløbet.

Til slut i forløbet er dygtigheden i 2017-analysen estimeret til 1,9 med en SEM på 0,51. Da den bedst mulige SEM der kunne opnås med 17 opgaver, er 0,48, kan det konstateres at det mindre end optimale valg af opgaver kun har haft en begrænset betydning for usikkerheden af denne elevs dygtighed. At en SEM-værdi på 0,51 må betragtes som et udtryk for en større usikkerhed end man bør forlange af en pædagogisk test, har ikke noget at gøre med udvalget af opgaver for denne elev. Det skyldes at antallet af opgaver eleven har besvaret, er for lille.

Oplysningerne der findes i tabellen, kan også bruges til at vurdere tilpasningen til Rasch-modellen ved beregninger af den betingede sandsynlighed for det observerede svarmønster givet at eleven har besvaret ni ud af 17 opgaver rigtigt. Det vil sige om det i det store og hele – på nær noget, der kan forklares som konsekvenser af ren tilfældighed – er sådan at eleven har svaret korrekt på de letteste opgaver og forkert på de vanskeligste.

I dette tilfælde er sandsynligheden for det observerede svarmønster 0,000093. Dette er en meget lille sandsynlighed, men hvis man undersøger samtlige mulige måder hvorpå der kan scores præcis ni point, vil man opdage at der er mange forløb der er endnu mere usandsynlige end det observerede forløb. Da den samlede sandsynlighed for at observere et svarmønster der er mere usandsynligt end det observerede, er 0,253, er konklusionen at forskellen på elevens svarmønster og Rasch-modellens forventninger til svarmønstret er insignifikant. Hypotesen om at det samlede svarmønster kommer fra en Rasch-model med sværhedsgraderne fra 2017, forkastes derfor ikke af denne test.

Med hensyn til spørgsmålet om hvorvidt der er svar på enkelte opgaver der ikke svarer til Rasch-modellens forventninger, afslører tabel 5.2 heller intet som kan betragtes som signifikant. Det korrekte svar på den næstsidste opgave er det mest usandsynlige, men sandsynligheden for at denne elev svarer korrekt på denne opgave, er trods alt lig med 0,24. Altså langt fra overraskende.

Da der heller ikke er noget, som kan betragtes som usandsynligt i forhold til det samlede resultatet for dele af testresultat, hvor det mest usandsynlige er at der scores syv point på de første 14 opgaver ($p = 0{,}16$), er der kun én konklusion: Dette testforløb er en smuk illustration af, hvordan et adaptivt testforløb forventes at forløbe. Med nogle få undtagelser til sidst i forløbet skiftes der mellem rigtige og forkerte svar, og det samlede antal korrekte svar er lig med halvdelen af det antal opgaver der er stillet. Testforløbet er desuden karakteriseret ved at de første tre opgaver passer til elevens niveau, og at der af den grund ikke spildes tid med for lette eller vanskelige opgaver i starten af forløbet.

Men sådan er det langt fra for alle elever alene af den grund at langt fra alle elever møder opgaver på deres eget niveau i de første opgaver i det adaptive forløb. Det ville klart være at foretrække hvis den adaptive algoritme var implementeret således at alle elever mødte opgaver tættere på deres faktiske dygtighed i starten af forløbet.

5.3.1 Analyser baseret på sværhedsgrader fra nationale test

Analyserne i det foregående eksempel illustrerer hvorledes den adaptive algoritme ville have fungeret hvis man havde benyttet sværhedsgraderne fra 2017 til beregningen af dygtigheden, men havde beholdt de udvalgte opgaver, selvom dette udvalg af opgaver havde forholdt sig til de historiske sværhedsgrader.

Venstre side af figur 5.1 og tallene i tabel 5.3 viser hvorledes tingene ville have fungeret hvis DNT's sværhedsgrader havde været korrekte.

Forskellen på forløbene i figur 5.1 er først og fremmest at det er tydeligt at udviklingen i sværhedsgrader i venstre side af figuren ligger tæt op ad udviklingen af estimaterne af dygtigheden således som den skal i et velfungerende adaptivt system.

Dette kan også ses hvis man sammenligner tallene i tabel 5.3 og tabel 5.2. I tabel 5.2 er forskellene mellem

Tabel 5.3 Oplysninger om testforløbet for elev 441985 baseret på DNT's sværhedsgrader.

		Betingede sandsynligheder				Dygtighedsestimater		
itemno	opgavenummer	score	scoreprob	cumulated	cumulatedprob	WML	SEM	Next
1	108020115072	1	0,730	1	0,730			
2	010802000301234810-1	0	0,417	1	0,487			
3	010802000301234966-1	1	0,725	2	0,492	2,69	1,06	2,79
4	010802000301234959-1	0	0,476	2	0,356	2,32	1,00	2,23
5	010802000301234953-1	1	0,668	3	0,403	2,65	0,91	2,65
6	010802000301234951-1	0	0,438	3	0,285	2,36	0,84	2,30
7	108020111026	1	0,650	4	0,350	2,61	0,78	2,70
8	0108020110139-1	0	0,450	4	0,239	2,40	0,72	2,48
9	108020111022	1	0,605	5	0,321	2,61	0,68	2,65
10	010802000301234848-1	0	0,437	5	0,202	2,43	0,64	2,62
11	108020111019	1	0,569	6	0,298	2,62	0,62	3,04
12	010802000301238021-1	1	0,457	7	0,405	2,82	0,60	2,92
13	108020111006	0	0,510	7	0,319	2,68	0,57	3,31
14	108020110268	0	0,612	7	0,196	2,59	0,55	3,46
15	0108020110166-1	1	0,350	8	0,436	2,77	0,53	3,72
16	010802000301239338-1	1	0,291	9	0,706	2,96	0,52	3,71
17	010802000301239337-1	0	0,706	9	1,000	2,88	0,50	

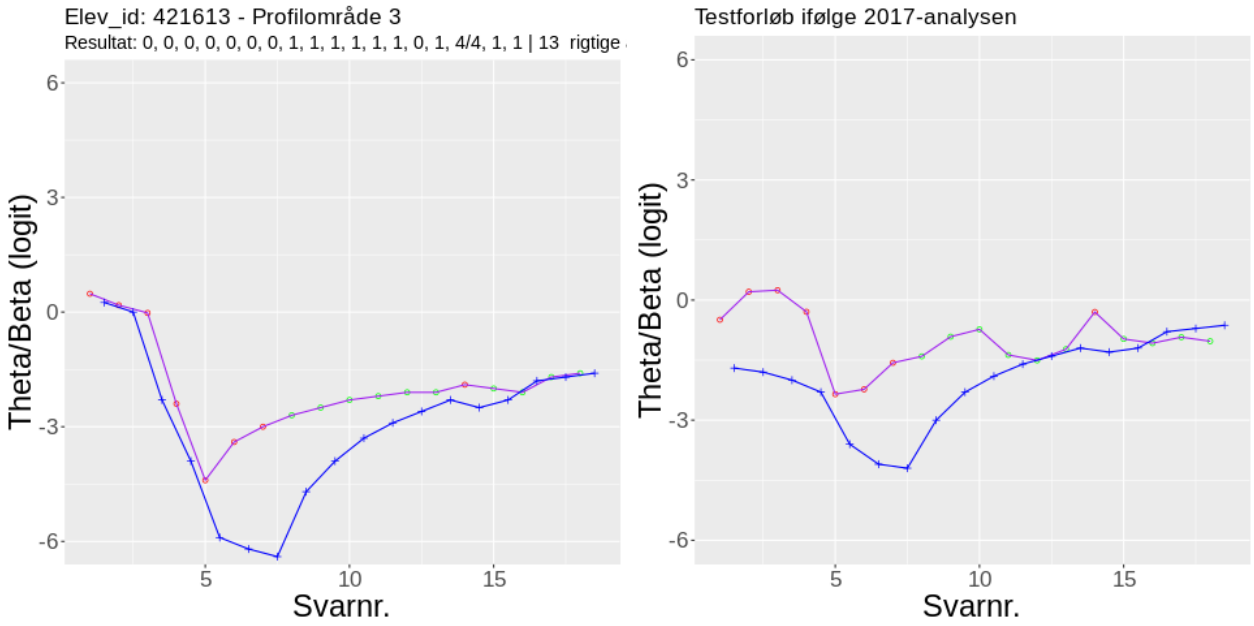
Note:
Se foregående tabel for forklaring af parametre.

de aktuelle estimater af dygtigheden og sværhedsgraden af den efterfølgende opgave mindre end i tabel 5.3. Da dygtighed og sværhedsgrad skal passe til hinanden, hvis den adaptive algoritme fungerer efter hensigten, skal der også være en stærk korrelation mellem dygtigheden og sværhedsgraden af den næste opgave. I tabel 5.3 er denne sammenhæng lig med 0,59, hvilket ganske vist er højsignifikant, men alligevel meget svagere end i tabel 5.2 hvor korrelationen er lig med 0,82.

Eller, med andre ord: hvis DNTs sværhedsgrader havde været anvendelige i 2017 ville den adaptive algoritmen have fungeret perfekt. Da det ikke er tilfældet, fungerer den dårligere.

5.4 Først flere forkerte svar, derefter mange rigtige

De følgende testforløb har alle først nogle forkerte og dernæst mange rigtige svar.



Figur 5.2 Testforløb for elev 421613, profilområde 3. Se forklaring i figur 5.1.

Elev 421613 lægger som det fremgår af figur 5.2, ud med syv forkerte. Eleven fortsætter med at lave fejl og algoritmen fortsætter med at reducere estimatet af dygtigheden indtil den efter syv forkerte er nede på -6,38 logit. Derefter laver eleven 13 korrekte og svarer kun et enkelt forkert.

Den ene af de 18 opgaver er en superitem-opgave med fire spørgsmål. I sådanne opgaver er sværhedsgraden ikke altid veldefineret. Nationale tests adaptive algoritme vælger derfor opgaver ud fra hvor opgaven er mest informativ. I dette tilfælde er det for personer med en dygtighed på -1.10. Omkring dette punkt er opgaven med de fire spørgsmål dobbelt så informativ som fire enkeltstående opgaver ville være. Hvis opgaven vælges på den rigtige måde vil testresultatet for denne elev altså svare til resultatet af 21 dikotome opgaver.

Figurerne afslører et fænomen som kan iagttages i flere testforløb fra 2017. Disse forløb er karakteriseret ved et påfaldende stort antal forkerte svar i begyndelsen af forløbet efterfulgt af perioder hvor eleverne tilsyneladende ikke har problemer med også at besvare vanskelige opgaver.

Visuelle indtryk som dem der ses i de to figurer, kan imidlertid være misvisende, blandt andet fordi den adaptive algoritme starter med at vælge opgaver i midten af fordelingen af dygtighederne, således at de mindst dygtige elever bliver bedt om at besvare opgaver der er alt for vanskelige for dem, og først får opgaver hvor der er en fair mulighed for et korrekt svar, når estimationen af dygtigheden kommer ned på elevens niveau.

I første omgang kunne man få det indtryk at dette kunne være tilfældet her. I hvert fald er sandsynligheden for det observerede forløb givet den samlede score på 13 ikke så usandsynligt så det i sig selv er udtryk for signifikant misforhold mellem det observerede og det forventede ($p = 0{,}24$).

Analyserne i tabel 5.4 fortæller imidlertid en anden historie. Beregningerne af sandsynlighederne for svar på enkelte opgaver finder i første omgang intet påfaldende. De mest usandsynlige begivenheder er de forkerte svar på den femte og sjette opgave hvor sandsynlighederne for at dette skulle ske er nede på henholdsvis 0,14 og 0,15. Men sandsynligheder af denne størrelsesorden kan ikke betragtes som udsagn om signifikante forskelle i forhold til det Rasch-modellen forventer. Sandsynligheden for at de syv første opgaver alle besvares forkert, er til gengæld højsignifikant med en p-værdi på 0,0001, især fordi opgaverne 5-7 alle er relativt lette Det indtryk man får i figur 5.2, er altså velbegrundet. Noget er gået galt i starten af forløbet, og da sværhedsgraderne jo er som de er, må det skyldes at elevens dygtighed ikke kommer til udtryk i svarene på de første opgaver.

Tabel 5.4 Oplysninger om testforløbet for elev 421613 baseret på 2017-sværhedsgraderne.

		Betingede sandsynligheder				Dygtighedsestimater		
itemno	opgavenummer	score	scoreprob	cumulated	cumulatedprob	WML	SEM	Next
1	01080306061340005-2	0	0,545	0	0,5449			
2	108030311018	0	0,705	0	0,3743			
3	01080306061330044-3	0	0,708	0	0,2529	-2,03	0,79	-0,26
4	010803060613252-4	0	0,599	0	0,1370	-2,33	0,74	-2,37
5	108030320029	0	0,142	0	0,0136	-3,59	0,86	-2,33
6	010803060613601-3	0	0,147	0	0,0010	-4,08	0,79	-1,59
7	1080306061330109	0	0,270	0	0,0001	-4,23	0,75	-1,44
8	0108030320022-3	1	0,698	1	0,0008	-3,00	0,90	-0,91
9	108030320040	1	0,569	2	0,0046	-2,35	0,84	-0,78
10	01080306912-1_2	1	0,537	3	0,0180	-1,89	0,75	-1,46
11	0108030612006-1	1	0,702	4	0,0366	-1,63	0,68	-1,54
12	108030320033	1	0,719	5	0,0635	-1,43	0,63	-1,22
13	1080306061330104	1	0,648	6	0,1135	-1,25	0,60	-0,35
14	0108030612019-1	0	0,576	6	0,0489	-1,33	0,58	-1,07
15	01080306061330034-2	1	0,610	7	0,1035	-1,17	0,55	-1,10
16	108030310373	4	0,405	11	0,3325	-0,79	0,45	-0,93
17	0108030320069-2	1	0,575	12	0,5997	-0,71	0,45	-1,03
18	01080303060940011-1	1	0,600	13	1,0000	-0,63	0,44	

Testforløbet for elev nr. 421613 afsluttes med at elevens dygtighed beregnes til -0,63 med en SEM på 0,44. SEM-værdien er mindre end den bedste SEM, som kunne opnås med 21 dikotome opgaver, hvilket skyldes den ekstra information som den polytome opgave bidrager med.

Vurderingen af elevens dygtighed må nødvendigvis undervurdere hvor dygtig eleven faktisk er, fordi beregningerne inddrager syv opgaver hvor elevens dygtighed ikke er kommet til udtryk. Wright og Stone (1979) som beskrev disse fænomener for næsten 40 år siden, foreslog derfor at dygtigheden skulle genberegnes fraregnet de syv opgaver. Gør man det, ændres estimatet af dygtigheden til 0,93 som er så meget større at det må konkluderes at det første bud på dygtigheden er misvisende i meget alvorlig grad. Konsekvensen er desværre også at SEM baseret på opgave 8-18 forøges til 0,79, og at konklusionen for denne elev må være at testresultatet ikke kan bruges. Estimatet baseret på alle 18 opgaver kan ikke bruges fordi resultatet er vildledende i svær grad, og resultatet baseret på opgave 8-18 kan ikke bruges fordi det er for usikkert.

5.5 Undersøgelse af udvalgte forløb

De analyser der er beskrevet i de to foregående afsnit, er foretaget for tilsammen 15 forskellige forløb. Detaljerne i analyserne kan ses i bilag B. Formålet med nærværende afsnit er at sammenfatte resultaterne af analyserne og at diskutere årsager til og konsekvenser af resultaterne.

5.5.1 Sammenfatning af resultater af analyser af enkeltstående forløb

Tabel 5.5 præsenterer en oversigt over test af tilpasning mellem enkeltstående testforløb og en Rasch-model med de parametre for sværhedsgrader som blev estimeret i 2017-analysen.

Tabel 5.5 Resultat af test af tilpasning mellem enkeltstående forløb og Rasch-modellen.

	Svarvektor		Mindste p	
Elev-id	Betinget sandsynlighed	p	Item	Delscore
104649	0,000	0,489	0,105	0,0063
129172	0,000	0,736	0,186	0,005
143590	0,003	0,838	0,237	0,0268
219768	0,001	0,351	0,110	0,0008
259724	0,024	0,956	0,307	0,0237
305503	0,021	0,760	0,196	0,0213
317854	0,000	0,397	0,187	0,0025
341070	0,002	0,901	0,340	0,1473
349294	0,000	0,984	0,230	0,0299
386356	0,000	0,082	0,077	0,0005
387213	0,111	0,793	0,322	0,1107
421613	0,000	0,241	0,142	0,0001
428314	0,001	0,961	0,171	0,059
439773	0,000	0,608	0,080	0,0137
441985	0,000	0,254	0,238	0,1567

Note:
Rækkerne er sorteret efter elev-id.
[1] Betinget sandsynlighed er den betingede sandsynlighed for svarmønstret givet den samlede score.
[2] P er p-værdien for den betingede sandsynlighed.
[3] Mindste p for items angiver p-værdien for det item der havde den laveste p-værdi.
[4] P-værdien for delscoren er den mindste p-værdi for den kumulerede score.

Tilpasningen til Rasch-modellen blev afvist i 11 ud af 15 tilfælde, i alle tilfælde fordi var tale om perioder med usandsynligt lave eller høje scores.

Tabel 5.6 viser en oversigt over de test vi har gennemført af tilpasning mellem enkeltstående forløb og Rasch-modellen. Tabellen gengiver estimater af elevdygtigheden (WML) og SEM for de enkelte forløb. I de tilfælde hvor tabel 5.5 afslørede problemer med enkelte dele af forløbet, oplyser tabel 5.6 også om estimater og SEM-værdier baseret på den del af forløbet der kan bruges (item-uddrag). Her er gengivet estimater af dygtigheden fraregnet opgaver med fejl i starten hvis antallet af sådanne fejl er signifikant. Endelig fortæller tabellen også om korrelationen mellem de løbende estimater af dygtigheden og den faktiske sværhedsgrad af den næste opgave.

Analysen afslørede fire forløb, hvor der ikke kunne påvises afvigelser mellem forløbene og Rasch-modellens forventninger, samt fire forskellige former for afvigelser fra det, som Rasch-modellen kan acceptere:

1) Seks forløb med for mange fejl i starten.
2) To forløb hvor eleven "står af" efter en i øvrigt god start.
3) To forløb hvor eleven står af efter en god start, men står på igen.
4) Et forløb hvor eleven tydeligt står af til sidst.

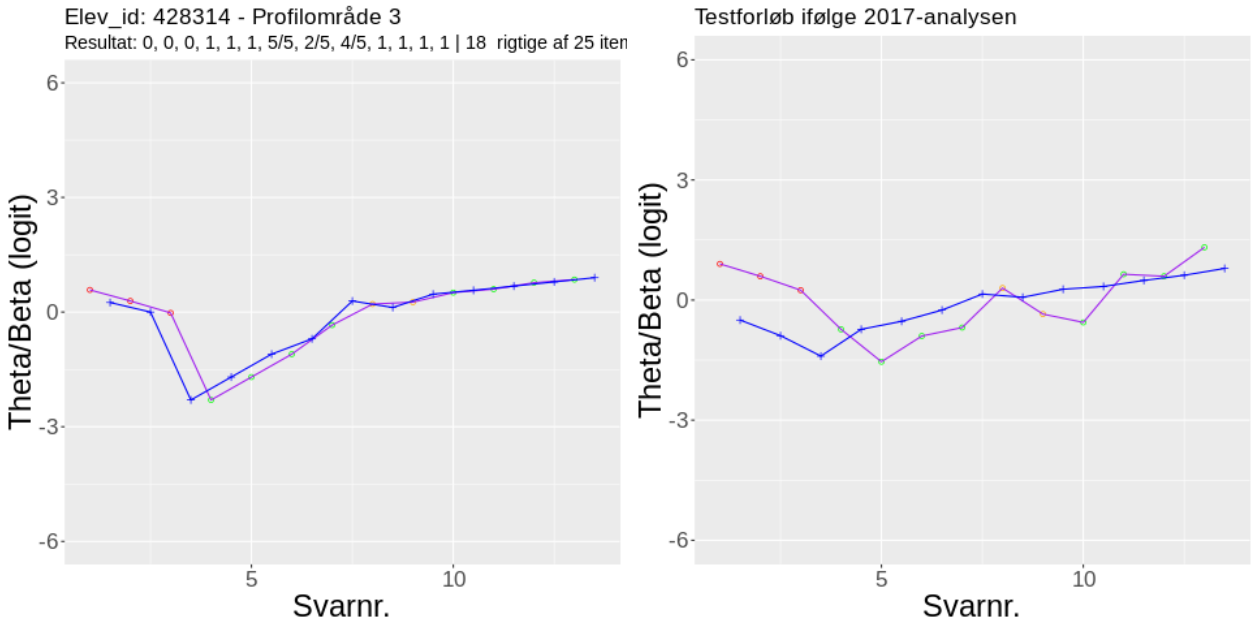
To af disse forløb er beskrevet i de foregående afsnit med detaljerede oplysninger om analyserne. De øvrige forløb vil blive beskrevet i det efterfølgende uden detaljerede analyseresultater. Interesserede læsere kan finde disse oplysninger i bilag A og B.

Tabel 5.6 Oversigt over test af tilpasning mellem enkeltstående forløb og Rasch-modellen.

			Alle items		Item-uddrag		
Elev_id	Score	Max	WML	SEM	WML	SEM	Person-item-korrelation
104649	15	25	-0,30	0,43	0,32	0,56	0,69
129172	38	61	0,46	0,28	0,70	0,30	0,73
143590	14	19	0,54	0,55	1,33	0,74	0,67
219768	5	15	2,10	0,61	-0,11	0,68	0,71
259724	6	10	0,08	0,69			0,67
305503	3	11	0,19	0,72	-1,62	0,71	0,79
317854	15	32	1,56	0,38			0,64
341070	9	17	2,36	0,54			0,65
349294	13	21	-0,33	0,44	0,20	0,56	0,71
386356	6	22	-0,71	0,52	-1,57	0,66	0,80
387213	4	10	2,03	0,76			0,86
421613	13	21	-0,63	0,44	0,93	0,79	0,64
428314	18	25	0,79	0,51			0,86
439773	9	23	1,70	0,47	1,17	0,56	0,78
441985	9	17	1,88	0,51			0,59

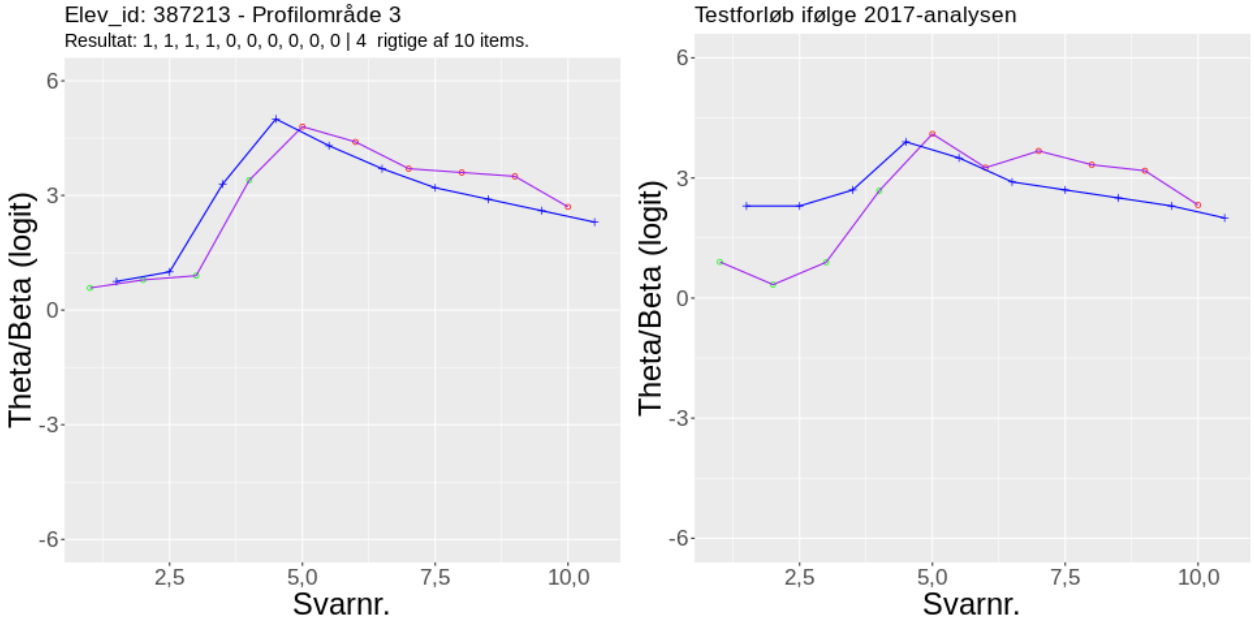
Note:
Rækkerne er sorteret efter elev-id.
[1] Score: Samlet score for eleven.
[2] Max: Den maksimale score i det givne forløb.
[3] Alle items, WML og SEM: Estimat og usikkerhed på estimatet af elevens dygtighed for hele forløbet.
[4] Item-uddrag, WML og SEM: Estimat og usikkerhed på estimatet af elevens dygtighed for den del af forløbet som ikke har udvist tegn på fejl.

5.5.2 Forløb, hvor der ikke kan påvises afvigelser fra Rasch-modellens forventninger



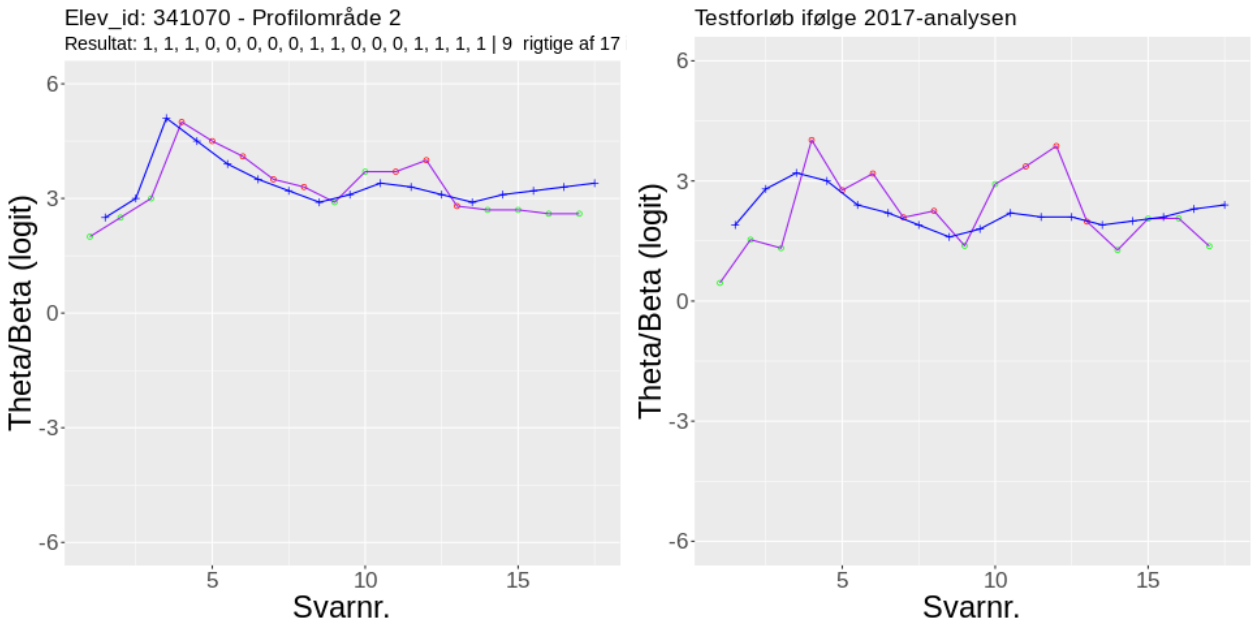
Figur 5.3 Testforløb for elev 428314, profilområde 3. Se forklaring i figur 5.1.

Elev 428314 svarer som det fremgår af figur 5.3, først tre forkerte, et antal som lige netop ikke er signifikant ($p = 0.05$), dernæst svarer hun rigtigt på 18 af 22 spørgsmål. Tre ud af 13 opgaver er polytome, hvilket bidrager til at få SEM ned på 0,51. Mon hun kunne have klaret udfordringen fra sværere spørgsmål og derved have fået et bedre resultat?



Figur 5.4 Testforløb for elev 387213, profilområde 3. Se forklaring i figur 5.1.

Som det fremgår af figur 5.4, lægger elev 387213 i tekstforståelsesprofilområdet ud med fire korrekt besvarede opgaver hvorefter hun tilsyneladende står af. I forhold til det afsluttende mål for dygtigheden er opgaverne lige i overkanten af hvad eleven burde klare. På trods af serien af forkerte svar ender hun omkring den 90. percentil. Der er derfor intet der kan tages som belæg for at eleven er stået af efter fjerde opgave. Da der kun er besvaret ti opgaver som rammer lidt ved siden af eleven, er SEM for denne elev på 0,76. Uanset at tilpasningen til Rasch-modellen ikke kan forkastes, burde dette testresultat være afvist. Det er simpelthen for usikkert. Denne elevs forløb er et eksempel på at der ingen lid kan fæstes til resultatet på grund af den meget høje SEM, og det adaptive system burde derfor have markeret over for læreren at dette resultat ikke bør bruges.

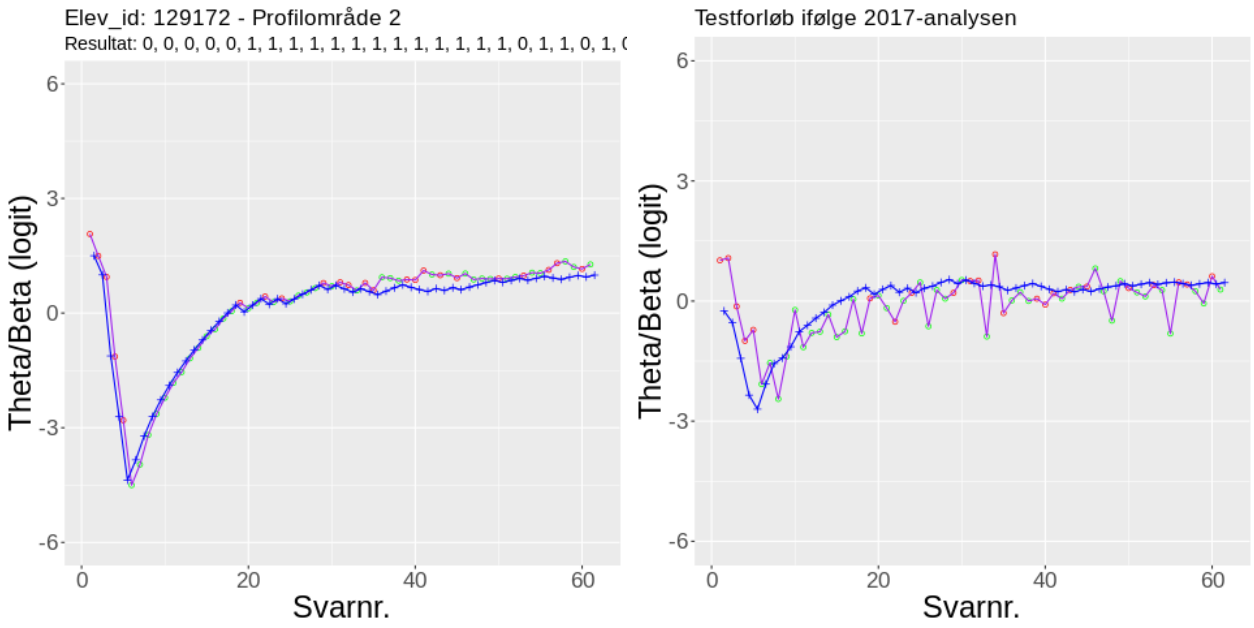


Figur 5.5 Testforløb for elev 341070, profilområde 2. Se forklaring i figur 5.1.

Elev 341070's forløb i profilområdet afkodning er karakteriseret ved klumper, se figur 5.5. Tre korrekte, fem fejl, to korrekte, tre fejl og fire korrekte. Det ser påfaldende ud, men tallene i sig selv kan ikke afvise tilpasning mellem responser og model. Sådan kan testresultater se ud, uden at der er noget galt.

Denne elevs forløb har præcis samme mønster som elev 317854, figur 5.11, i de første 17 opgaver, men hvor elev 341070 stopper efter 17 svar, får 317854 får lov at fortsætte med yderligere 15 spørgsmål. Det resulterer i helt forskellige resultater for de to elever.

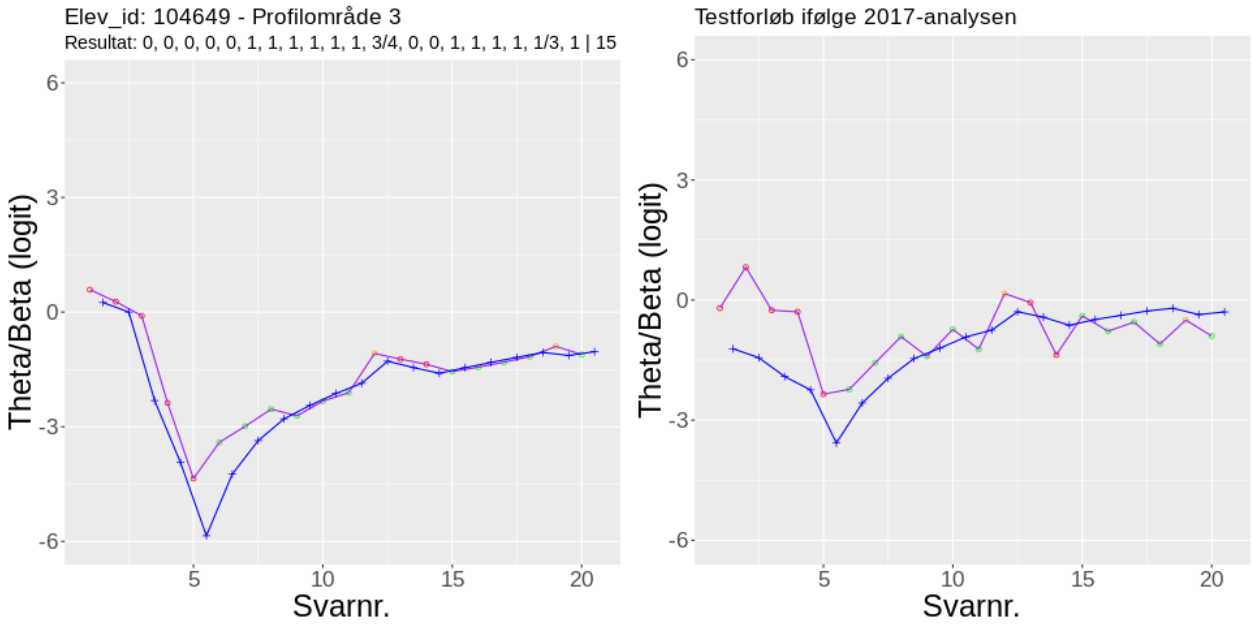
5.5.3 Forløb med dårlig start



Figur 5.6 Testforløb for elev 129172, profilområde 2. Se forklaring i figur 5.1.

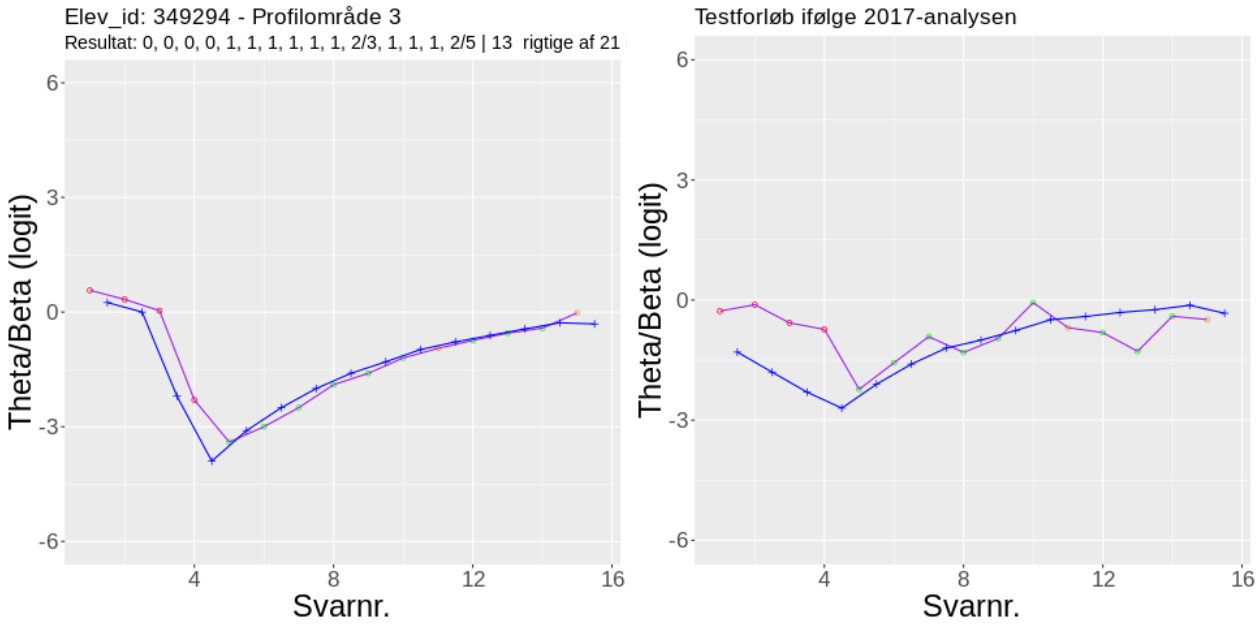
Elev 129172 lægger ud med 5 forkerte svar, heraf tre som eleven ikke burde have vanskeligheder med, se figur 5.6. Derefter ser forløbet ud til at passe meget flot til Rasch-modellen. Da der besvares 61 opgaver, kommer SEM ned på 0,27. Hvis man udelader de første fem opgaver forbedres estimatet af dygtigheden fra 0,46 til 0,70 med en SEM på 0,30.

Dette er et eksempel på et forløb med tilstrækkelig mange opgaver til at producere relativt sikre målinger, og hvor en justering der korrigerer for den dårlige start, ikke forringer målingerne. Det er faktisk det eneste af de 15 eksempler hvor vi kan konkludere på denne måde.



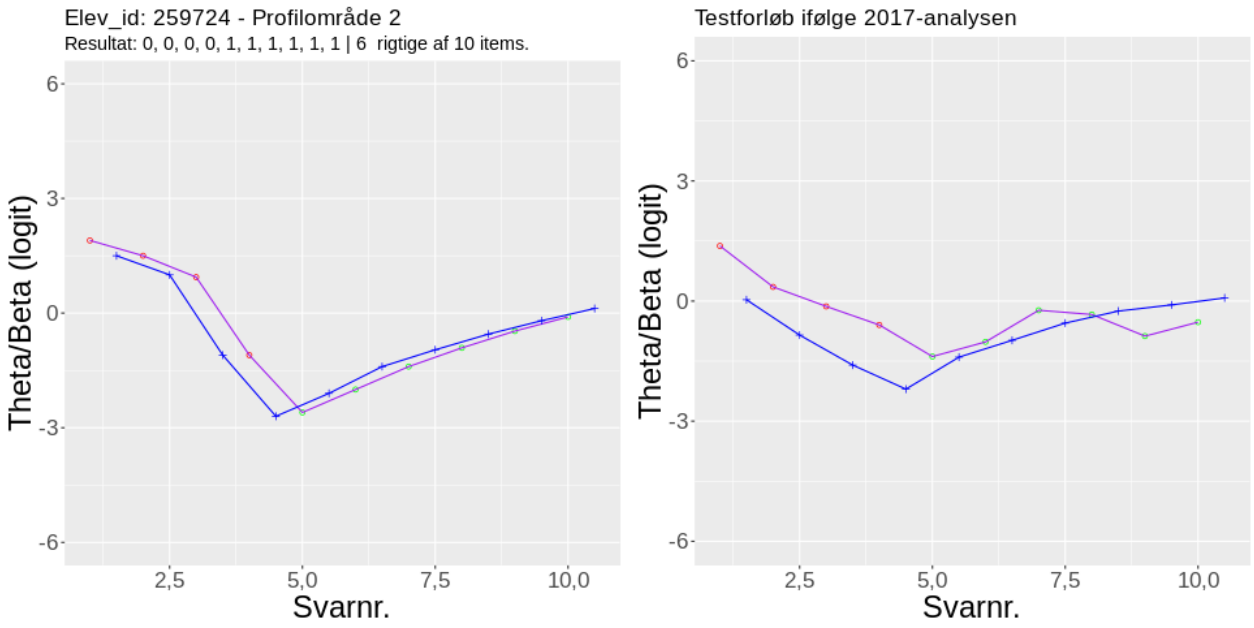
Figur 5.7 Testforløb for elev 104649, profilområde 3. Se forklaring i figur 5.1.

Elev 104649, figur 5.7, lægger også ud med fem forkerte svar, dernæst ni rigtige (heraf tre i et partial credit-item)), tre forkerte (heraf et i et partial credit item) og fem rigtige (ét i et partial credit-item). Kurven flader ud undervejs i den lange række af rigtige. En hypotese kunne være at nationale tests algoritme på grund af et for lavt estimat stiller for lette opgaver og derved ikke giver eleven mulighed for at opveje den uheldige start. Hvis de første fem opgaver udelades af beregningerne, stiger dygtigheden fra -0,30 til 0,32 og SEM fra 0,43 til 0,56.



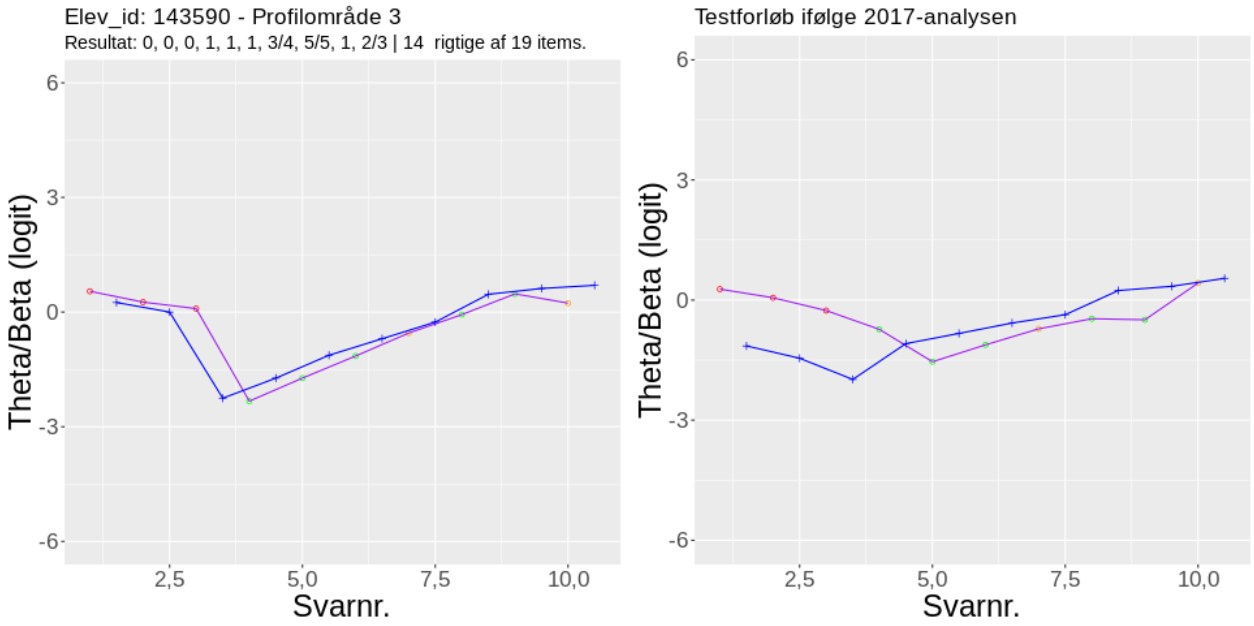
Figur 5.8 Testforløb for elev 349294, profilområde 3. Se forklaring i figur 5.1.

Elev 349294 lægger ud med fire forkerte svar, se figur 5.8. Derefter svarer hun rigtigt på 13 af 17 items. Afvigelsen fra Rasch-modellen er kun svagt signifikant (p = 0,030). Hvis de første fire opgaver udelades af beregningerne, stiger dygtigheden fra -0,33 til 0,20 og SEM fra 0,44 til 0,56. Ville hun have fortsat den lange stribe af korrekte svar hvis hun havde fået lov at fortsætte?



Figur 5.9 Testforløb for elev 259724, profilområde 2

Elev 259724, se figur 5.9, er meget længe om at svare på hver enkelt opgave, se tabel A.9. Dette forløb starter også med fire fejl, men er herefter fejlfrit. Ville hun have fortsat sin stribe af rigtige hvis hun havde fået lov at fortsætte? Afvigelsen fra Rasch-modellen er kun svagt signifikant (p = 0,023). Der er kun besvaret 10 opgaver, som – hvis de alle inkluderes – giver en SEM på 0,69. Dette testresultat bør forkastes med og uden de indledende opgaver. Usikkerheden er så stor, at resultatet er ubrugeligt.



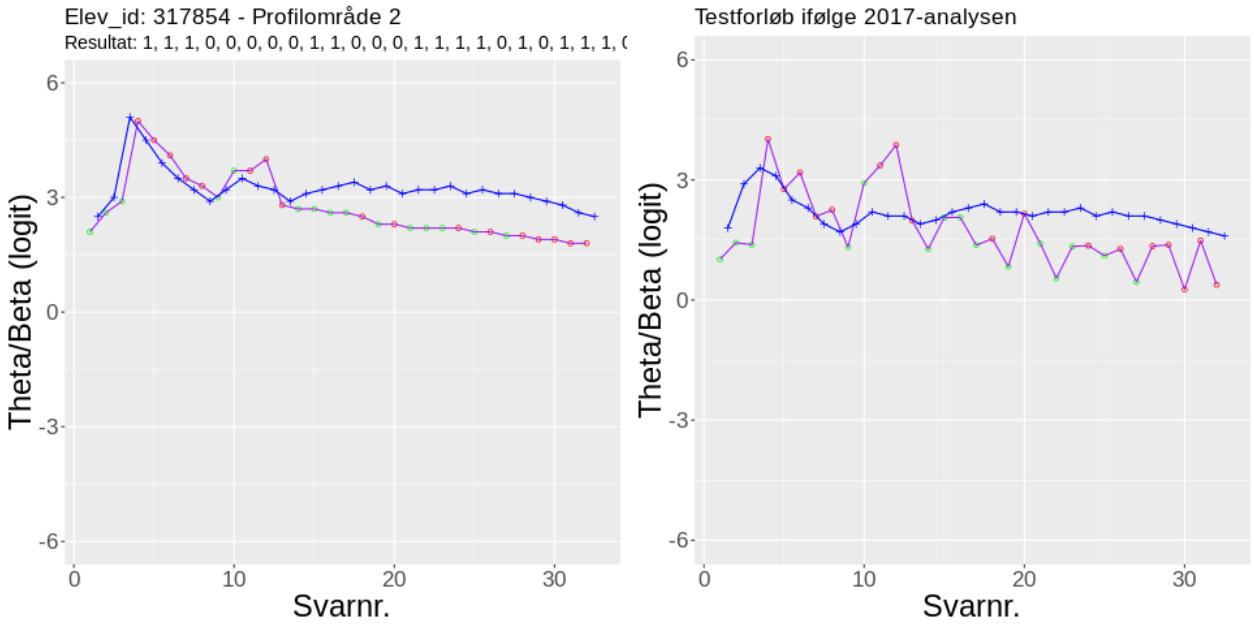
Figur 5.10 Testforløb for elev 143590, profilområde 3. Se forklaring i figur 5.1.

Elev 143590 starter sit forløb i tekstforståelse med tre fejl, se figur 5.10. Dernæst svarer hun rigtigt på 14 og forkert på to delspørgsmål. På trods af næsten tre gange så mange rigtige som forkerte, når hun kun lige op over hvor hun startede. Problemet med dette forløb er at eleven indledningsvist svarer forkert på opgaver som hun ifølge det endelige estimat burde have kunnet svare rigtigt på, men da hun så begynder at svare rigtigt, er opgaverne for lette, og hun kan ikke trække sit resultat langt nok op efter ti opgaver.

Ville eleven kunne have fortsat sin lange række af rigtige svar hvis hun havde fået mere tid? Afvigelsen fra Rasch-modellen er kun svagt signifikant (p = 0,027). Hvis de første tre opgaver udelades af beregningerne, stiger dygtigheden fra 0,54 til 1,33 og SEM fra 0,55 til 0,74 hvilket er langt fra det tilladelige. Dette forløb bør også forkastes selvom man kan korrigere for den usikre start.

5.5.4 Forløb med dårlig slutning

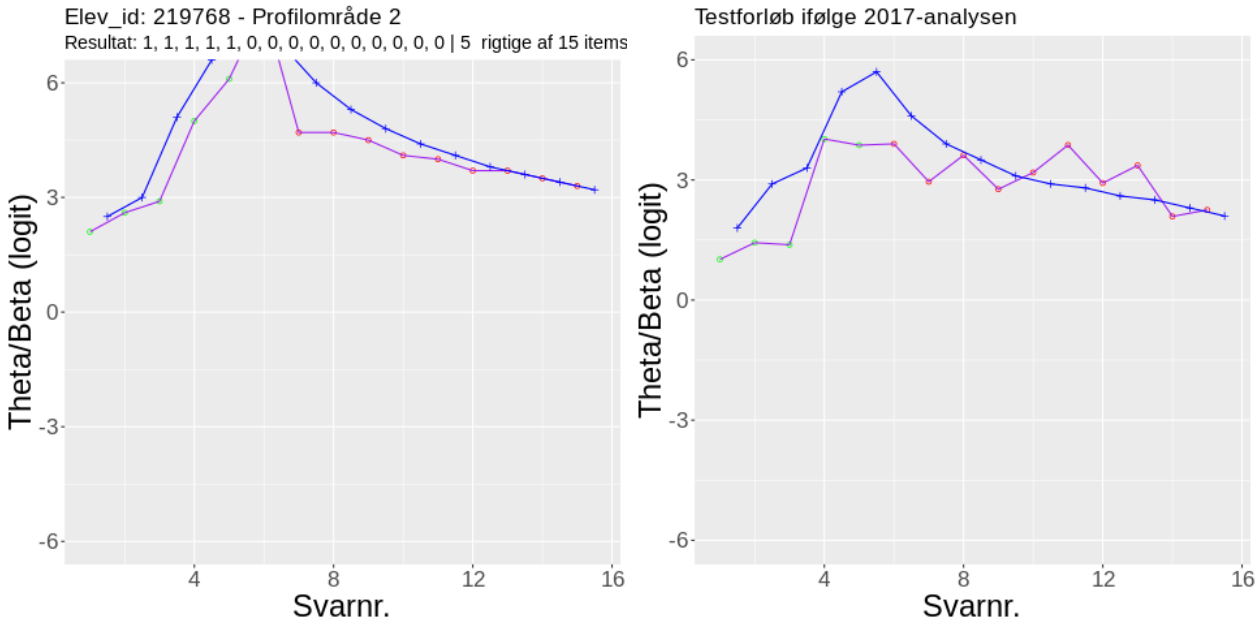
De følgende forløb stammer fra elever som starter med flere korrekte svar, men dernæst går over til at svare forkert.



Figur 5.11 Testforløb for elev 317854, profilområde 2. Se forklaring i figur 5.1.

Eleven 317854, se figur 5.11, har i de første 17 opgaver i profilområdet afkodning nøjagtig det samme svarmønster som elev 341070, se figur 5.5, men får lov at fortsætte med yderligere 15 spørgsmål. Det resulterer i et væsentligt dårligere resulat omkring den 55. percentil.

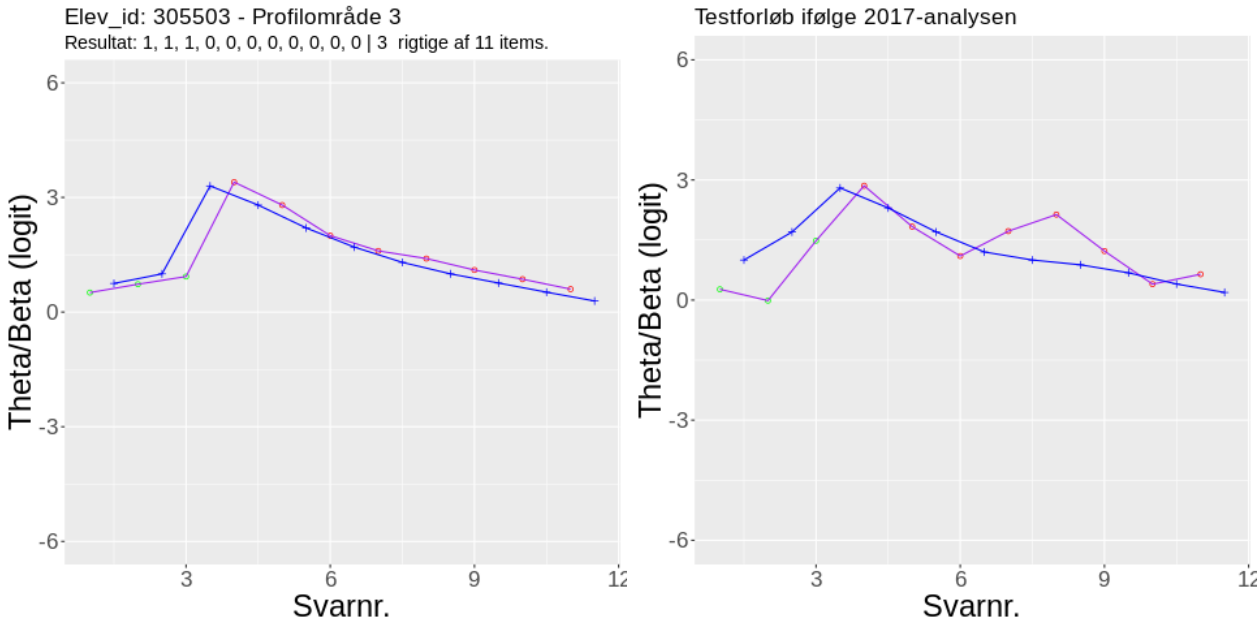
Eleven slutter af med fem fejl på opgaver hvor eleven har mere end en fair chance for at svare rigtigt. Undervejs er der dog andre variationer med klumper af svar der passer ret dårligt til sværhedsgraderne. Vores vurdering er at der enten er tale om en ujævn præstation hvor indsatsen fra elevens side varierer, eller hvor der er typer af opgaver som eleven har problemer med. Uden adgang til indholdet af opgaverne og uden mulighed for at tale med eleven kan vi blot konkludere at slutresultatet formodentlig undervurderer elevens dygtighed. Bemærk i tabel A.11, at eleven har stoppet forløbet undervejs og genoptaget det seks dage senere.



Figur 5.12 Testforløb for elev 219768, profilområde 2. Se forklaring i figur 5.1.

Elev 219768, se figur 5.12, lægger ud med at svare rigtigt på fem spørgsmål. De resterende ti svarer hun forkert på. De mange korrekte betyder naturligvis at den adaptive algoritme herefter vælger meget vanskelige opgaver, men dette er ikke nok til at forklare at der ikke forekommer en eneste korrekt opgave i de sidste ti opgaver. Forløbet afviger højsignifikant fra Rasch-modellens forventninger, og fem rigtige ud af 15 opgaver i en adaptiv test fortæller i sig selv at resultatet ikke kan bruges.

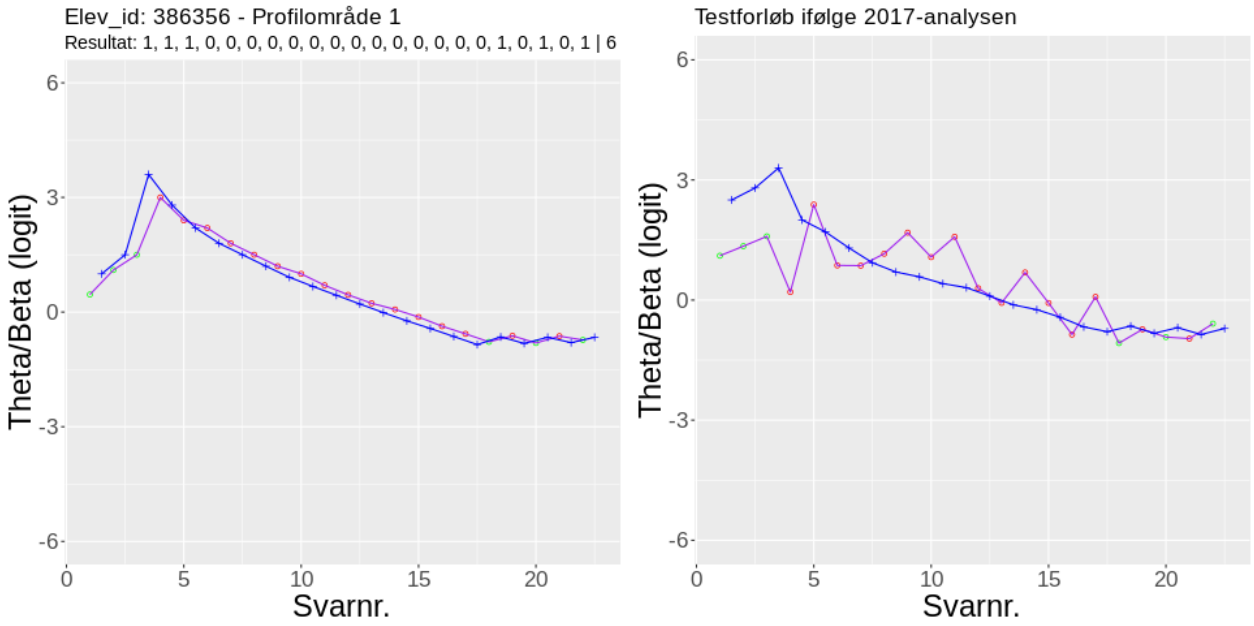
Hvad ville der ske hvis eleven havde fået yderligere 10 spørgsmål, måske nogle lettere spørgsmål? Ville hun også have svaret forkert på dem? Dette er et eksempel på hvor vigtigt det er at svare rigtigt på de første spørgsmål.



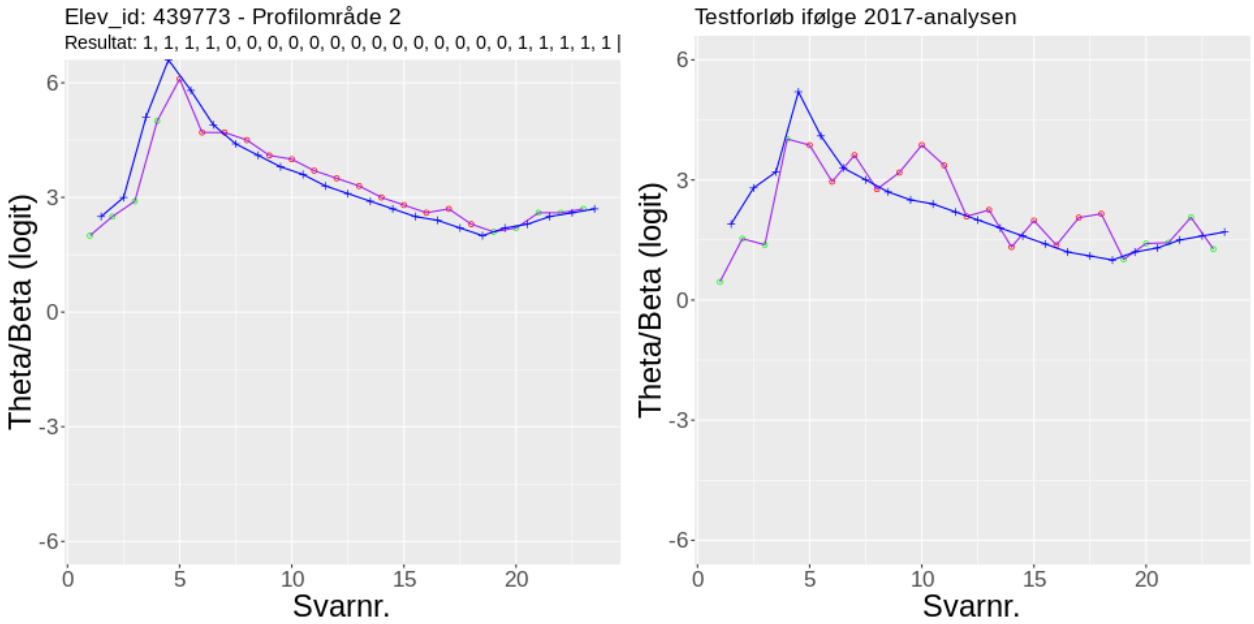
Figur 5.13 Testforløb for elev 305503, profilområde 3. Se forklaring i figur 5.1.

Elev 305503 svarer rigtigt på tre spørgsmål og dernæst forkert på otte hvorefter hun stoppes, se figur 5.13. De indledende korrekte svar betyder naturligvis at den adaptive algoritme herefter vælger meget vanskelige opgaver, men dette er ikke helt nok til at forklare at der ikke forekommer en eneste korrekt opgave i de sidste otte opgaver. Forløbet afviger svagt signifikant fra Rasch-modellens forventninger, og tre rigtige ud af 11 opgaver i en adaptiv test signalerer også at resultatet ikke kan bruges. Beregningerne slutter i øvrigt med en SEM på 0,72. Alene af den grund burde resultatet være forkastet.

5.5.5 Forløb med rigtige, så forkerte og så rigtige igen



Figur 5.14 Testforløb for elev 386356, profilområde 1. Se forklaring i figur 5.1.



Figur 5.15 Testforløb for elev 439773, profilområde 2. Se forklaring i figur 5.1.

Elev 386356, se figur 5.14, indleder i profilområdet sprogforståelse med tre korrekte svar og svarer dernæst forkert på 14 spørgsmål og så rigtigt på tre spørgsmål og forkert på fem spørgsmål. Af naturlige grunde vælger det adaptive system fem meget lette opgaver til sidst, men det er ikke nok til at bortforklare at

Tabel 5.7 Percentil- og logitplacering af de undersøgte elevers resultater.

		Percentiler		Logit	
Elev_id	Profilområde	DNT-analysen	2017-analysen	DNT-analysen	2017-analysen
104649	3	5%	16%	-1,037	-0,297
129172	2	15%	19%	0,993	0,460
143590	3	42%	45%	0,700	0,543
219768	2	80%	77%	3,195	2,099
259724	2	6%	11%	0,115	0,078
305503	3	24%	32%	0,286	0,192
317854	2	56%	57%	2,549	1,556
341070	2	85%	85%	3,392	2,361
349294	3	10%	15%	-0,310	-0,334
386356	1	19%	23%	-0,661	-0,708
387213	3	88%	91%	2,321	2,027
421613	3	4%	9%	-1,556	-0,631
428314	3	50%	54%	0,896	0,788
439773	2	63%	62%	2,731	1,698
441985	2	70%	69%	2,891	1,883

Note:
Percentiler er for begge analysers vedkommende beregnet på baggrund af fordelingen af elevdygtighederne i 2017 (og er altså ikke de historiske percentiler som DNT rapporterer til lærerne).

der er noget der ikke passer. I lyset af at eleven laver så mange fejl er tre rigtige på de første tre opgaver klart signifikant (p = 0.00048). Seks korrekte ud af 22 opgaver er også alt for lidt i en adaptiv test selvom nationale tests algoritmes valg af nye opgaver ikke svarer perfekt til opgavernes sande sværhedsgrader. Dette testresultat er klart uanvendeligt. Bemærk at de 6 sidste spørgsmål i testen kun er stillet inden for dette profilområde, se tabel A.14, så algoritmen har ikke været sikker på estimatet før den 22. opgave i dette profilområde.

Fire rigtige svar i profilområdet afkodning bringer elev 439773 meget højt op, se figur 5.15. De følgende 11-12 forkerte får hende tilbage hvor hun startede. En slutspurt med fem rigtige til sidst opvejer de foregående fem forkerte svar. Eleven har siddet ved testen i to timer og et kvarter. Variationerne kan igen til dels forklares ved at der kommer mange vanskelige opgaver efter den gode start og mange lette opgaver til sidst på grund af de mange fejl, men variationerne er svagt signifikante (p = 0,013).

5.6 Afsluttende kommentarer til analyserne

Målet med en adaptiv algoritme er at give eleven opgaver som passer til elevens niveau. Man har valgt at lade algorimten starte med at give opgaver på samme niveau til alle elever, og derfor må man forvente at mange elever svarer forkert eller rigtigt på et antal opgaver i starten, indtil algoritmen har fundet et passende estimat af elevens dygtighed. Det er svært at afgøre præcis hvor mange opgaver det vil være rimeligt at forvente at et større antal elever svarer forkert/rigtigt på i træk. I tabel 5.8 har vi angivet hvor mange elever der starter med at svare konsekvent rigtigt eller forkert på op til ti opgaver i træk.

Knap 7 procent af alle testforløb (9.509 testforløb (hver elev gennemfører tre testforløb)) indeholder enten fire forkerte svar eller fire korrekte svar på opgaver i træk i starten af forløbet. Et ikke ubetydeligt antal

Tabel 5.8 Antal elever der svarer konsekvent rigtigt eller forkert i starten af forløbet

	Svarer forkert			Svarer korrekt		
Antal fejl/korrekt i træk	Profilområde 1	Profilområde 2	Profilområde 3	Profilområde 1	Profilområde 2	Profilområde 3
2	13989	9689	7652	8052	15124	18262
3	6548	3855	3254	3575	9541	10634
4	1179	848	1275	2419	1449	2339
5	199	262	193	865	413	262
6	94	85	94	626	179	38
7	49	31	64	534	151	11
8	40	27	55	473	104	7
9	38	24	51	241	84	1
10	35	21	47	224	69	1

Note:
Tallene angiver hvor mange elever der har svaret korrekt på antallet af opgaver i træk. Tallene indeholder også elever der har svaret mere end det givne antal korrekt.

elever svarer også forkert eller rigtigt på fem eller flere opgaver i træk i starten af et af profilområderne. Det er med andre ord store andele af eleverne som ikke oplever en adaptiv test i væsentlige dele af forløbet.

Gennemgangen af testforløb har også vist at det at acceptere en SEM på 0,55 er alt for usikkert, og under alle omstændigheder bør ingen forløb med endnu højere SEM rapporteres til lærere, forældre eller andre (det fremgik i kapitel 4 at omkring 12-16 procent af forløbene resulterer i en SEM højere end 0,55 i alle tre profilområder).

Det vil altid være tilfældet i pædagogiske test at der er elever som ikke svarer i overensstemmelse med hvad man kan forvente givet den estimerede dygtighed. En rimelig konsekvens af dette vil være at elever som udviser uventet testadfærd, identificeres, og at det meddeles til læreren at denne elevs resultat ikke kan anvendes i det videre arbejde med eleven. Vi har med ovenstående analyser vist at sådanne forløb kan identificeres, og at man derfor kan gøre noget ved det (fx se bort fra opgaver eller forkaste resultatet).

Men det skal også understreges at dette problem ikke er et særtræk ved det adaptive system, men et generelt problem for alle test. It-baserede test (adaptive eller lineære) kan opdage det mens det sker, hvilket man således kunne udnytte, fx til at gøre læreren opmærksom på problemet undervejs.

6

Konklusioner og anbefalinger

Analyserne i denne rapport har afdækket en række bekymrende træk ved de nationale test. En opsamling på analyserne findes i det indledende resume, kapitel (2.3). Der er samlet set tale om:

a) at itemsværhedsgraderne som nationale tests algoritme bruger, ikke stemmer med de oplevede sværhedsgrader i 2017,
b) at dygtighederne for nogle elevers vedkommende estimeres forkert,
c) at nationale test for mange elevers vedkommende ikke måler så præcist som lovet,
d) at en for stor andel af elevernes testforløb ikke stemmer med forudsætningerne i Rasch-modellen, samt
e) at algoritmen udviser tegn på stiafhængighed (eller stopper for tidligt i processen).

I forhold til den sidste konklusion, er spørgsmålet om man kan gøre noget på forhånd. Det vil blandt andet kræve noget af læreren som jo har ansvaret for at instruere eleverne på den bedste måde så de ikke falder ud undervejs eller til sidst eller tror at de skal skynde sig for at besvare så mange opgaver som muligt.

Vi vil opfordre til at man i adaptive systemer tager højde for at der kan være gået noget galt i starten. Hvis eleverne af den ene eller den anden grund er nervøse, ukoncentrerede eller bliver overraskede over opgaverne i starten af forløbet og bliver slået ud af at de får uløselige opgaver, så ødelægger det testresultaterne. Et forslag kunne være at man lader læreren bestemme hvor systemet skal starte (enten for klassen eller for enkelte elever).

En anden løsning kunne være at blande de meget lette/svære spørgmål som er resultatet af mange indledende forkerte/rigtige, med spørgsmål af middel sværhedsgrad for at sikre at elevens indledende præstation ikke skyldtes held, utilpashed eller andre forstyrrelser.

En tredje løsning kunne være kun at anvende opgaver fra midterfeltet i run in-perioden så det indledende estimat af elevdygtigheden ikke ender så ekstremt som det gør ved tre indledende ens svarresultater.

Alternativt kunne man lade systemet starte med tre opgaver der er så lette at der ikke er nogen elever der får problemer, men som heller ikke bidrager med noget. Denne løsning har klare fordele, men den koster rent tidsmæssigt og forøger SEM'en.

Begrebet *stiafhængighed* bruges i samfundsvidenskaben til at beskrive det fænomen at når et politisk system er valgt, bliver det vanskeligere at skifte til et andet system. Selvom beregning af dygtigheden i Rasch-modeller er noget helt andet, og selvom resultaterne af forkerte svar på slutresultatet, hvis eleven kommer til at svare forkert på lette opgaver, er de samme uanset om det sker i starten eller i midten eller i slutningen af resultatet, så kan situationer, hvor starten af testforløbet af den ene eller anden grund mislykkes, føre til forløb der synes at være præget af stiafhængighed.

Dette skyldes to forhold. For det første at fire fejl i de første opgaver, af årsager der intet har med elevens færdigheder at gøre, under alle omstændigheder vil trække den samlede vurdering af dygtigheden ned. Det vil se slemt ud i starten, men efterhånden som eleven svarer på nye opgaver på en måde der (kun) afhænger af hvor dygtig eleven er, vil det samlede estimat blive realistisk om end en smule for lavt på grund af starten. Præcis hvor mange ekstra opgaver der skal til, er så vidt vi ved ikke undersøgt, men der er ingen grund til at tro at forløb med 10-20 opgaver kan kompensere for en mislykket start med fire eller flere fejlslagne svar på opgaver.

Dette problem er gældende for alle almindelige test. Det er kun fordi der er tale om adaptive test, at problemerne bliver synlige.

Den anden årsag er at der er tale om en adaptiv test. Den adaptive algoritme vil får det indtryk, at der er tale om en meget svag elev og vil derfor vælge opgaver, der er alt for lette for eleven. De korrekte svar vil naturligvis trække vurderingen af dygtigheden op, men da der er tale om lette opgaver, vil det ikke være lige så stærkt som hvis eleven havde vist at hun kunne svare korrekt på vanskelige opgaver. Til syvende og sidst vil vurderingen også blive realistisk, men der skal mange flere opgaver til end i almindelige test, fordi den adaptive algoritme vil forsøge at "vise hensyn" og ikke give eleven opgaver over det alt for lave niveau som algoritmen har placeret eleven på på grund af den mislykkede start på forløbet.

I den forstand har alle pædagogiske test problemer der minder om stiafhængighed, og problemerne er særlig mærkbare, når det drejer sig om adaptive test.

De ret bekymrende resultater der er opregnet først i kapitlet, giver i øvrigt anledning til flere overvejelser.

For det første er de problemer der er identificeret i denne rapport, ikke fuldstændig overraskende. I 2015 opdagede matematiklærer Jørgen Damgaard fra Ørum Skole i Norddjurs at der var meget stor forskel på resultater fra to nationale test taget med få dages mellemrum. Dette førte til at Norddjurs kommune stillede Undervisningsministeriet en række spørgsmål (Harbo 2015) som siden ledte til at Undervisningsministeriet udgav et notat om resultaterne i den pågældende klasse (Styrelsen for Undervisning og Kvalitet 2016). I forbindelse hermed udarbejdede ministeriet også et overordnet notat om "Undersøgelse af de nationale tests reliabilitet" (Styrelsen for It og Læring 2016b). Dette notat indeholder en række iagttagelser som burde have fået alarmklokkerne til at lyde.

I figur 6.1 ses resultaterne af en korrelationsanalyse af sammenhængen mellem resultater for elever der har taget to på hinanden følgende frivillige test i efteråret 2014. 32.566 elever tog den samme test to gange, og der var i gennemsnit 22 dage imellem testene og maksimalt 53 dage. Korrelationskoefficienterne er et udtryk for hvor stor overensstemmelse der er imellem de to resultater. Fuldstændig overensstemmelse (alle er blevet lige meget dygtigere mellem de to test) vil give værdien 1. I rapporten gives følgende retningslinjer for vurderingen af koefficienterne:

> Generel guideline til vurdering af reliabilitet: '0,0-0,5'=uacceptabel; '0,5-0,6'=dårlig; '0,6-0,7'=tvivlsom; '0,7-0,8'=acceptabel; '0,8-0,9'=god; '0,9-1,0'=fremragende

Disse tommelfingerregler stammer tilsyneladende fra generelle retningslinjer om korrelationer. Men i dette tilfælde er der tale om test-gentest-korrelationer som er et blandt flere mål for reliabiliteten af test (Feinman 2008). Som omtalt tidligere stilles normalt krav om en realibilitet på mindst 0,9 for individuelle målinger af elever.

Tabel 6 Sammenhængen mellem forsøg 1 og forsøg 2 i elevdygtigheden målt på logit skalaen. Pearson korrelationen

Test	Antal elever	Profil-område 1	Profil-område 2	Profil-område 3
Dansk/læsning 2. klasse	6.057	0,57*	**0,80***	**0,75***
Dansk/læsning 4. klasse	4.421	0,63*	**0,78***	**0,75***
Dansk/læsning 6. klasse	5.134	0,56*	**0,81***	**0,73***
Dansk/læsning 8. klasse	2.558	0,66*	**0,74***	**0,74***
Matematik 3. klasse	7.590	0,60*	0,56*	0,66*
Matematik 6. klasse	5.440	0,61*	0,56*	0,60*
Engelsk 7. klasse	2.424	**0,76***	**0,77***	**0,80***
Fysik/kemi 8. klasse	1.169	0,45*	0,41*	0,39*
Biologi 8. klasse	761	0,41*	0,49*	0,50*
Geografi 8. klasse	1.028	0,47*	0,47*	0,45*

* Statistisk signifikant forskellig fra 0

Figur 6.1 Tabel 6 i Styrelsen for It og Læring 2016a: Korrelationer mellem elevernes resultater ved en frivillig og en obligatorisk test.

Som det fremgår af tabellen lever otte af de i alt 30 profilområder op til karakteristikken "acceptabel", og tre opnår karakteristikken "god". De 19 øvrige profilområder udviser tegn på "tvivlsom", "dårlig" eller sågar "uacceptabel" korrelation mellem de to tests.

I notatets sammenfatning af undersøgelsen (s. 1) omtales dette resultat således:

> I alle profilområder er der en statistisk signifikant positiv sammenhæng mellem elevdygtigheden bestemt ved første og ved andet forsøg i de frivillige test. Specielt i afkodning og tekstforståelse i dansk læsning samt i engelsk er der en høj korrelation mellem to gentagne test.
>
> Generelt er korrelationen mellem elevens samlede vurdering i forsøg 1 og forsøg 2 på 0,79.

I vores øjne er denne sammenfatning ikke på nogen måde et rimeligt udtryk for de problemer som afdækkes med tabellen over korrelationer. Denne tabel tyder på grundlæggende problemer med testen, og det forekommer os uforståeligt at den ikke gav anledning til grundige undersøgelser af om der var fejl i nationale tests algoritme eller i estimeringen af itemsværhedsgrader.

I notatet beregnes også om udviklingen fra test 1 til test 2 fordeler sig som ventet. Notatet forklarer fremgangsmåden således:

> For hver elev beregnes en standardiseret forskel, U, på den estimerede elevdygtighed
>
> $$U = \frac{D_1 - D_2}{\sqrt{(SEM_1^2 + SEM_2^2)}}$$
>
> Her er D_1 og D_2 lig den estimerede elevdygtighed til første og andet forsøg, mens SEM_1 og SEM_2 er de tilhørende estimerede usikkerheder. Justeres endvidere med den gennemsnitlige niveauforskel fra første til andet forsøg, da vil U følge en standard normalfordeling. I denne fordeling forventes 95 % af elevernes resultater at ligge mellem -1,96 og +1,96 mens 5 % af elevernes resultater forventes at ligge udenfor ±1,96 (Styrelsen for It og Læring 2016b, 8).

Resultaterne af denne analyse fremgår af figur 6.2.

Tabel 5 Andelen af elever, hvor forskellen i elevdygtigheden i første og andet forsøg ligger uden for 95 % sikkerhedsinterval

Test	Antal elever	Profil-område 1	Profil-område 2	Profil-område 3
Dansk/læsning 2. klasse	6.057	16 %	20 %	24 %
Dansk/læsning 4. klasse	4.421	13 %	11 %	12 %
Dansk/læsning 6. klasse	5.134	11 %	8 %	11 %
Dansk/læsning 8. klasse	2.558	14 %	14 %	8 %
Matematik 3. klasse	7.590	12 %	11 %	14 %
Matematik 6. klasse	5.440	13 %	9 %	12 %
Engelsk 7. klasse	2.424	5 %	12 %	9 %
Fysik/kemi 8. klasse	1.169	8 %	6 %	9 %
Biologi 8. klasse	761	8 %	5 %	6 %
Geografi 8. klasse	1.028	7 %	7 %	8 %

Figur 6.2 Tabel 5 i Styrelsen for It og Læring 2016a: Andelen af elever hvor forskellen mellem estimaterne af dygtighed ligger uden for 95 procent-sikkerhedsintervallet.

Som det fremgår af tabellen, opfylder i bedste fald to af de 30 profilområder kravet om at være under 5 procent. De resterende 28 har værdier der er større. I profilområde 3 for læsning i 2. klasse har ikke mindre end 24 procent af eleverne fået en score ved den anden test der adskiller sig signifikant fra den forventede.

Tabellen kommenteres således:

> I alt ligger 12 procent af forskellene mellem elevdygtighederne i forsøg 1 og forsøg 2 udenfor det forventede, hvilket er lidt mere end de 5 procent, der forventes i en normalfordeling (tabel 5). Spredningen i elevernes resultater mellem første og andet forsøg er således lidt større end, der forventes, og større end usikkerheden (SEM) på elevdygtighederne kan forklare (Styrelsen for It og Læring 2016b, 8).

Karakteristikken af forskellen mellem de forventede (og acceptable) fem procent og de iagttagede 12 procent som "lidt større end der forventes" finder vi upræcis. Der er tale om en forskel på 140 procent. Og også denne iagttagelse burde have ført til undren og yderligere undersøgelser. Dette resultat omtales ikke i sammenfatningen til notatet.

Notatet formidler flere resultater end disse, fx at de opgaver eleverne får ved den anden test, for 25 procents vedkommende allerede er set ved den første test. For dansk, læsning 8. klasse drejer det sig om 50 procent af opgaverne.

Som sagt er det vores opfattelse af disse resultater hver for sig og samlet burde have ført til en meget grundig undersøgelse af måleegenskaberne ved nationale test i stil med den undersøgelse vi har gennemført i denne rapport.

6.1 Usikkerheden på resultaterne

Der har som omtalt i indledningen været stort fokus på de problemer den store usikkerhed på resultaterne fra nationale test indebærer. Denne undersøgelse har vist at der for ganske store andele af elever er tale om endog større usikkerheder end de i forvejen alt for høje målestandardfejl på 0,55 logit. I praksis betyder det at læreren ikke får informationer om sine elever som hun ikke allerede havde en god fornemmelse af. En pædagogisk test bør have en SEM på mindre end 0,3 logit sådan som det var forventningen i de første år af testens levetid.

6.2 Konsekvenser for forskning, kvalitetssikring og beslutninger om indsatser

Vi har i denne rapport peget på en række bekymrende træk ved nationale tests måleegenskaber. Disse resultater må føre til at aktørerne på en lang række områder tager hidtidige beslutninger op til fornyet overvejelse.

6.2.1 Forskning baseret på nationale test – kan vi stole på den?

Der er foretaget en del forskning der har taget udgangspunkt i resultater fra nationale test. Det handler fx om undersøgelser af sammenhænge mellem elevbaggrund og nationale test, det handler om vurdering af udviklings- og forskningsindsatser osv. Denne undersøgelse må resultere i at resultater fra forskning og evaluering der har taget udgangspunkt i data fra nationale test, tages op til fornyet undersøgelse.

6.2.2 Skolereformen – er succeskriteriet troværdigt?

Nationale test blev ved vedtagelsen af skolereformen i 2013 indskrevet i den politiske aftale som et blandt flere måltal:

> Mål 1. Folkeskolen skal udfordre alle elever, så de bliver så dygtige, de kan
>
> Resultatmål 1.1. for mål 1: Mindst 80 pct. af eleverne skal være gode til at læse og regne i de nationale test ("Aftale Mellem Regeringen (Socialdemokraterne, Radikale Venstre Og Socialistisk Folkeparti), Venstre Og Dansk Folkeparti Om et Fagligt Løft Af Folkeskolen" 2013).

Nærværende undersøgelse rejser tvivl om hvorvidt der kan fæstes lid til de hidtidige evalueringer af dette måltal.

6.2.3 Beslutninger om lærere og skoler – kan vi stole på dem?

Nationale test blev oprindeligt præsenteret som et pædagogisk redskab (Bundsgaard 2018c; Undervisningsminister Bertel Haarder (V) 2006), men blev kort efter vedtagelsen også integreret i lov nr. 572 om ændring af lov om folkeskolen. Heri hed det blandt andet at:

> »§ 13 a. Undervisningsministeren offentliggør på baggrund af de afholdte test, jf. § 13, stk. 3, hvert år landsresultaterne.
>
> Stk. 2. Kommunalbestyrelsen og skolelederen får oplyst den enkelte skoles testresultater set i forhold til det samlede landsresultat. Heri skal indgå en vurdering af skolens testresultater set i forhold til elevernes sociale baggrund.« ("Lov Om Ændring Af Lov Om Folkeskolen" 2006)

Nærværende undersøgelse af nationale tests måleegenskaber må føre til overvejelser over om den kommunale og lokale kvalitetssikring af skoler og undervisning har hvilet på et forkert grundlag.

6.2.4 Beslutninger om elever – var de velbegrundede?

Nationale test skal indgå i lærernes arbejde med eleverne, og forældrene skal informeres om resultatet. Derfor har nationale test opnået en central placering i beslutningsprocesserne om eleverne (Bundsgaard og Puck 2016). Resultaterne er indgået i beslutninger om tildeling af specialindsatser, ved fastsættelse af standpunktskarakterer, ved holdinddeling af elever, ved beslutning om behov for ekstra fokus på faglige områder osv.

Denne undersøgelse har sandsynliggjort at alle disse beslutninger hviler på et usikkert grundlag og derfor bør genevalueres.

6.3 Pædagogiske test i skolen

Begrundelsen ved folketingsdebatten i 2005-2006 for indførelsen af nationale test var at testene skulle fungere som et pædagogisk redskab. Der er i vores øjne ikke tvivl om at pædagogiske test kan tjene meget vigtige formål både i forhold til diagnostisering af elever med faglige vanskeligheder, som led i lærerens evaluering af sin undervisning og forberedelse til ny undervisning, som indsigt i faglige kompetencer og udfordringer for grupper af elever mv. Nationale test kan ikke fungere som et sådant redskab i den nuværende version.

Vores opfordring er på denne baggrund at nationale test sættes på pause indtil såvel årsagerne som konsekvenserne af problemerne er grundigt undersøgt. Vores opfordring er også at man genovervejer hvilke værktøjer lærerne faktisk har brug for for at skabe den forbedrerede evalueringskultur der var begrundelsen for at indføre nationale test.

Vi vil opfordre til at man løsriver brugen af test som pædagogisk redskab fra brugen af test til kvalitetssikring.

Og vi vil opfordre til at man udvikler en række af kvalitetstest som lærerne kan tage i anvendelse når de ser et behov for det.

Sådanne test skal for det første teste et meget bredere spektrum af de faglige mål end nationale test gør.

For det andet skal de bygge på nyere forskning i autentiske testsituationer (fx Bundsgaard 2018a; Bundsgaard 2016; Care, Scoular, og Griffin 2016; Siddiq, Gochyyev, og Wilson 2017; Geisinger 2016; Greenstein 2012; Griffin og Care 2015).

Og for det tredje skal de have tilstrækkelig høj psykometrisk kvalitet.

På trods af den indlysende fordel ved adaptive test i forhold til at opnå større sikkerhed i resultatet på kortere tid, er vores vurdering på baggrund af denne rapport og vores tidligere arbejde med nationale test, at den helt igennem adaptive tilgang som er valgt med nationale test, resulterer i større problemer end den løser. Vores anbefaling er derfor at man anvender mindre omfattende adaptive tilgange, som fx det såkaldte *staged adaptive testing* hvor eleverne svarer på en række opgaver hvorefter de tildeles en ny række opgaver som passer til det niveau besvarelsen af den første sekvens af opgaver har estimeret dem til. Derved kan man udvikle mere avancerede opgaver, som kan teste mere varierede aspekter af de faglige områder.

Litteratur

Adams, Raymond J., og Mark Wilson. 1996. "Formulating the Rasch Model as a Mixed Coefficients Multinomial Logit Model". I *Objective Measurement: Theory into Practice Volume 3*, 143–66. Norwood, NJ: Ablex Publishing.

Adams, Raymond J., og Margaret L. Wu. 2007. "The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model". I *Multivariate and Mixture Distribution Rasch Models*, redigeret af Matthias von Davier og Claus H. Carstensen, 57–76. New York: Springer Science Business Media.

Adams, Raymond J., Mark Wilson, og Wen-chung Wang. 1997. "The Multidimensional Random Coefficients Multinomial Logit Model". *Applied Psychological Measurement* 21 (1): 1–23.

"Aftale Mellem Regeringen (Socialdemokraterne, Radikale Venstre Og Socialistisk Folkeparti), Venstre Og Dansk Folkeparti Om et Fagligt Løft Af Folkeskolen". 2013.

Boch, R.D., og M. Aitken. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm". *Psykometrika*, nr. 46: 443–59.

Bundsgaard, Jeppe. 2016. "Rapport Om Kompetencetest i Hitte På-Projektet". København: DPU, Aarhus Universitet.

———. 2018a. "Det 21. Århundredes Kompetencer". I *Skoleudvikling Med It: Forskning i Tre Demonstrationsskoleforsøg I*, redigeret af Jeppe Bundsgaard, Marianne Georgsen, Stefan Graf, og Thomas Illum Hansen, 143–65. Aarhus: Aarhus Universitetsforlag.

———. 2018b. "Test Og Måling i Fagene". I *Udvikling i Didaktik Didaktik i Udvikling*, redigeret af Torben Spanget Christensen, Nikolaj Elf, Peter Hobel, og Ane Qvortrup. Aarhus: Klim.

———. 2018c. "Pædagogisk brug af test". *Sakprosa* 10 (2). doi:10.5617/sakprosa.6007.

Bundsgaard, Jeppe, og Morten Rasmus Puck. 2016. *Nationale Test - Danske Lærere Og Skolelederes Brug, Holdninger Og Viden*. DPU, Aarhus Universitet.

Care, Esther, Claire Scoular, og Patrick Griffin. 2016. "Assessment of Collaborative Problem Solving in Education Environments". *Applied Measurement in Education* 29 (4): 250–64. doi:10.1080/08957347.2016.1209204.

Feinman, Joshua. 2008. "High Stakes, but Low Validity? A Case Study of Standardized Tests and Admissions

into New York City Specialized High Schools." Boulder; Tempe: Education and the Public Interest Center & Education Policy Research Unit.

Geisinger, Kurt F. 2016. "21st Century Skills: What Are They and How Do We Assess Them?" *Applied Measurement in Education* 29 (4): 245–49. doi:10.1080/08957347.2016.1209207.

Greenstein, Laura M. 2012. *Assessing 21st Century Skills: A Guide to Evaluating Mastery and Authentic Learning*. Corwin Press.

Griffin, Patrick, og Esther Care, red. 2015. *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-9395-7.

Hale, Charles Dennis, og Douglas Astolfi. 2014. *Measuring Learning and Performance: A Primer*. 3rd edition. St. Leo, Florida: Saint Leo University.

Harbo, Ulf. 2015. "Norddjurs kommune udfordrer de Nationale test - Folkeskolen.dk". https://www.folkeskolen.dk/573670/norddjurs-kommune-udfordrer-de-nationale-test.

Hoover, Wesley A., og Philip B. Gough. 1990. "The Simple View of Reading". *Reading and writing* 2 (2): 127–60.

Kousholt, Kristine. 2015. "Børns Gætterier Ved Nationale Test". *Cepra-striben*, Nationale Tests„ nr. 18: 46–57.

Kreiner, Svend. 2007. "Den Adaptive Procedure".

Kreiner, Svend, og Karl Bang Christensen. 2013. "Person Parameter Estimation and Measurement in Rasch Models". I *Rasch Models in Health*, redigeret af Karl Bang Christensen, Svend Kreiner, og M Mesbah, 63–78. London: ISTE & John Wiley & Sons.

"Lov Om Ændring Af Lov Om Folkeskolen". 2006.

Norling, Marina. 2016. "De nationale test 2016 - hvor galt står det til? (1) - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/586828/de-nationale-test-2016—hvor-galt-staar-det-til-1.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in Mathematical Psychology, Vol. 1. Copenhagen: Danmarks Pædagogiske Institut.

Ravn, Karen. 2014. "Ups de nationale test måler ikke så præcist som lovet - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/539694/ups–de-nationale-test-maaler-ikke-saa-praecist-som-lovet.

———. 2015a. "Duer ikke: En femtedel af opgaverne i de nationale test kasseret - Folkeskolen.dk". *folkeskolen.dk*. https://www.folkeskolen.dk/572751/duer-ikke-en-femtedel-af-opgaverne-i-de-nationale-test-kasseret.

———. 2015b. "Eksperter dumper de nationale test - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/572813/eksperter-dumper-de-nationale-test.

———. 2015c. "Skoleleder: Testresultater svinger, som vinden blæser - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/572808/skoleleder-testresultater-svinger-som-vinden-blaeser.

Riise, Andreas Brøns. 2014. "Nationale test: Klasse havde kæmpe udsving på tre dage - Folkeskolen.dk". https://www.folkeskolen.dk/540229/nationale-test-klasse-havde-kaempe-udsving-paa-tre-dage.

Robitzsch, Alexander, Thomas Kiefer, og Margaret Wu. 2019. *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM.

Rosenbaum, Paul R. 1989. "Criterion-Related Construct Validity". *Psychometrika* 54 (4): 625–33. doi:10.1007/BF02296400.

Siddiq, Fazilat, Perman Gochyyev, og Mark Wilson. 2017. "Learning in Digital Networks Literacy:

A Novel Assessment of Students' 21st Century Skills". *Computers & Education* 109 (juni): 11–37. doi:10.1016/j.compedu.2017.01.014.

Smarter Balanced Assessment Consortium. 2018. "Smarter Balanced Assessment Consortium: 2016-17 Technical Report".

Styrelsen for It og Læring. 2015. "Den adaptive algoritme i De Nationale Test". København: Undervisningsministeriet, Styrelsen for It og Læring.

———. 2016a. *Test- Og Prøvesystemet - De Nationale Test. Brugervejledning for Skoler*. København: Ministeriet for Børn, Unge og Undervisning.

———. 2016b. "Undersøgelse af de nationale tests reliabilitet". København: Undervisningsministeriet.

———. 2016c. "Nationale Tests Måleegenskaber". København: Undervisningsministeriet.

———. 2017. "Vejledning Til Nye Resultatvisninger i de Nationale Test Til Lærere i Alle Fag". København: Styrelsen for It og Læring.

Styrelsen for Undervisning og Kvalitet. 2016. "Undersøgelse Af Udsving i Testresultater På Ørum Skole i Norddjurs Kommune". Styrelsen for Undervisning og Kvalitet.

Taylor, Wilson L. 1953. "Cloze Procedure: A New Tool for Measuring Readability". *Journalism & Mass Communication Quarterly* 30 (4): 415–33.

Undervisningsminister Bertel Haarder (V). 2006. "L 101 Forslag til lov om ændring af lov om folkeskolen. (Styrket evaluering og anvendelse af nationale test som pædagogisk redskab samt obligatoriske prøver m.v.)."

Wandall, Jakob. 2010. "Test, prøver og evaluering i grundskolen". Powerpoints. København: IND/KU.

Wandall, Jakob, Christine Nørrelund, og Mette Dalgaard Nielsen. 2018. "Elevernes Syn På de Nationale Test". Nordic Metrics.

Wells, Craig S, og James A Wollack. 2003. "An Instructor's Guide to Understanding Test Reliability". Testing & Evaluation Services. University of Wisconsin.

Wilson, Mark. 2003. "On Choosing a Model for Measuring". *Methods of Psychological Research-Online* 8 (3): 1–22.

———. 2005. *Constructing Measures : An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Wright, Benjamin D., og Mark H. Stone. 1979. *Best Test Design*. Chicago: MESA Press.

Wu, Margaret, Raymond J. Adams, Mark Wilson, og S. Haldane. 2007. *ConQuest: Generalised Item Response Modelling Software (Version 2.0)*. Camberwell, Australia: ACER Press.

Zwinderman, A. H. 1995. "Pairwise Parameter Estimation in Rasch Models". *Applied Psychological Measurement*, nr. 19: 369–75.

A

Testforløbstabeller

I disse tabeller kan man se

1) hvornår opgaven er besvaret,
2) hvor mange delspørgsmål eleven har haft korrekt,
3) hvor mange delspørgsmål der var i opgaven, samt
4) hvad algoritmen estimerede elevens dygtighed (theta) til efter besvarelsen af opgaven. Bemærk at dette ikke gælder de to første spørgsmål, idet de tre første spørgsmål fungerer som "run in", dvs. at systemet starter med en opgave på et forvalgt sværhedsgradsniveau og derfra stiller et væsentligt sværere spørgsmål (0,25 eller 0,5 logit sværere) hvis eleven svarer rigtigt – og lettere hvis eleven svarer forkert. Efter det tredje spørgsmål estimeres elevens dygtighed for første gang.
5) Desuden fremgår det hvilket nummer spørgsmål dette er for eleven i alt,
6) hvilket nummer spørgsmål det er inden for profilområdet,
7) tiden det tog for eleven i minutter at besvare opgaven (beregnet ud fra starttidspunkt på denne opgave og den følgende).
8) I den ottende søjle angives 2017-analysens estimat af elevens dygtighed (theta-2017). Estimatet er foretaget allerede efter besvarelsen af den første opgave.
9) For 2017-analysen er også angivet usikkerheden på estimatet (SEM-2017).
10) Beta-DNT er nationale tests estimat af opgavens sværhedsgrad, og
11) Beta-2017 er opgavens sværhedsgrad beregnet med udgangspunkt i eleverne fra 2017.

Tabel A.1: Testforløbdata for elev 441985, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020115072	1	1	2,5	2	1	0,51	2,7	2,20	1,9	1,377
010802000301234810-1	0	1	2,0	5	2	1,50	1,4	1,40	2,6	1,433
010802000301234966-1	1	1	2,9	8	3	1,30	1,9	1,20	2,0	1,346
010802000301234959-1	0	1	2,3	11	4	1,60	1,6	1,00	2,8	1,985
010802000301234953-1	1	1	2,7	14	5	0,50	1,9	0,91	2,2	1,411
010802000301234951-1	0	1	2,4	17	6	0,68	1,6	0,83	2,7	2,058
0108020111026	1	1	2,6	20	7	0,43	1,8	0,77	2,3	0,837
0108020110139-1	0	1	2,4	23	8	0,91	1,5	0,72	2,7	1,274
0108020111022	1	1	2,6	26	9	0,98	1,7	0,68	2,5	1,532
010802000301234848-1	0	1	2,4	29	10	2,60	1,5	0,64	2,6	2,066
0108020111019	1	1	2,6	31	11	0,82	1,7	0,61	2,6	1,370
010802000301238021-1	1	1	2,8	33	12	0,66	1,8	0,59	3,0	1,323
0108020111006	0	1	2,7	35	13	1,60	1,7	0,56	2,9	1,381
0108020110268	0	1	2,6	37	14	0,73	1,6	0,54	3,3	2,251
0108020110166-1	1	1	2,8	39	15	2,10	1,7	0,53	3,5	2,089
010802000301239338-1	1	1	3,0	42	16	0,28	1,9	0,52	3,7	2,922
010802000301239337-1	0	1	2,9	45	17	0,57	1,9	0,51	3,7	3,363

Tabel A.2: Testforløbdata for elev 421613, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010803060613400005-2	0	1	0,25	3	1	0,56	-1,70	2,20	0,480	-0,493
0108030311018	0	1	0,00	6	2	1,90	-1,80	1,90	0,180	0,206

Tabel A.2: Testforløbdata for elev 421613, profilområde 3 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010803060613300044-3	0	1	-2,30	9	3	0,84	-2,00	1,80	-0,023	0,246
010803060613252-4	0	1	-3,90	12	4	1,10	-2,30	1,70	-2,400	-0,294
0108030320029	0	1	-5,90	15	5	0,86	-3,60	1,80	-4,400	-2,351
010803060613601-3	0	1	-6,20	18	6	1,90	-4,10	1,70	-3,400	-2,229
01080306061330109	0	1	-6,40	21	7	0,48	-4,20	1,70	-3,000	-1,567
0108030320022-3	1	1	-4,70	24	8	1,30	-3,00	1,00	-2,700	-1,406
0108030320040	1	1	-3,90	27	9	0,90	-2,30	0,84	-2,500	-0,915
01080306912-1__2	1	1	-3,30	30	10	2,00	-1,90	0,74	-2,300	-0,731
0108030612006-1	1	1	-2,90	33	11	2,10	-1,60	0,67	-2,200	-1,371
0108030320033	1	1	-2,60	36	12	0,84	-1,40	0,63	-2,100	-1,506
01080306061330104	1	1	-2,30	39	13	1,70	-1,20	0,60	-2,100	-1,228
0108030612019-1	0	1	-2,50	42	14	2,10	-1,30	0,58	-1,900	-0,298
01080306061330034-2	1	1	-2,30	45	15	0,85	-1,20	0,55	-2,000	-0,972
0108030310373	4	4	-1,80	48	16	0,78	-0,79	0,44	-2,100	-1,077
0108030320069-2	1	1	-1,70	51	17	1,90	-0,71	0,44	-1,700	-0,927
01080303060940011-1	1	1	-1,60	54	18	0,83	-0,63	0,44	-1,600	-1,028

Tabel A.3: Testforløbdata for elev 428314, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010803000301238841-1	0	1	0,25	3	1	2,60	-0,500	2,10	0,580	0,903
0108030320023-2	0	1	0,00	6	2	1,00	-0,890	1,90	0,290	0,594
010803060613300044-3	0	1	-2,30	9	3	0,74	-1,400	1,80	-0,023	0,246
01080306912-1__2	1	1	-1,70	12	4	1,30	-0,730	1,10	-2,300	-0,731
0108030310024-3	1	1	-1,10	15	5	0,77	-0,530	1,00	-1,700	-1,541
01080306061330114	1	1	-0,70	18	6	1,20	-0,250	0,89	-1,100	-0,898
0108030310611-2	5	5	0,29	21	7	1,90	0,150	0,70	-0,340	-0,687
010803000301239063-1	2	5	0,12	24	8	4,10	0,069	0,56	0,210	0,300
010803000301241409-1	4	5	0,47	39	9	3,20	0,270	0,51	0,260	-0,351
0108030310606-1	1	1	0,57	42	10	1,10	0,340	0,50	0,510	-0,559
010803000301237450-3	1	1	0,68	45	11	1,90	0,490	0,50	0,600	0,645
010803000301235608-1	1	1	0,79	48	12	3,90	0,620	0,50	0,770	0,594
010803000301238350-1	1	1	0,90	51	13	1,40	0,790	0,50	0,850	1,317

Tabel A.4: Testforløbdata for elev 387213, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010803000301238841-1	1	1	0,75	3	1	2,6	2,3	2,10	0,58	0,903
0108030320062-1	1	1	1,00	6	2	1,4	2,3	1,90	0,79	0,335
0108030320053-1	1	1	3,30	9	3	1,0	2,7	1,80	0,90	0,895
010803000301235582-2	1	1	5,00	12	4	3,5	3,9	1,90	3,40	2,680
010803000301235587-2	0	1	4,30	15	5	5,0	3,5	1,30	4,80	4,100
010803000301235606-2	0	1	3,70	18	6	4,6	2,9	1,00	4,40	3,256
010803000301235578-2	0	1	3,20	21	7	2,6	2,7	0,93	3,70	3,674
010803000301236038-1	0	1	2,90	24	8	4,4	2,5	0,86	3,60	3,329
010803000301235602-2	0	1	2,60	27	9	3,2	2,3	0,80	3,50	3,180
010803000301235608-2	0	1	2,30	30	10	3,4	2,0	0,76	2,70	2,323

Tabel A.5: Testforløbdata for elev 341070, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020111016	1	1	2,5	2	1	0,32	1,9	2,10	2,0	0,452
0108020111022	1	1	3,0	5	2	0,65	2,8	2,00	2,5	1,532
010802000301238021-1	1	1	5,1	8	3	0,68	3,2	1,80	3,0	1,323
010802000301239468-1	0	1	4,5	11	4	1,00	3,0	1,40	5,0	4,022
0108020110231	0	1	3,9	14	5	1,10	2,4	1,10	4,5	2,768

Tabel A.5: Testforløbdata for elev 341070, profilområde 2 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010802000301234818-1	0	1	3,5	17	6	0,20	2,2	0,95	4,1	3,185
0108020110166-1	0	1	3,2	20	7	0,86	1,9	0,86	3,5	2,089
0108020110268	0	1	2,9	23	8	0,28	1,6	0,81	3,3	2,251
0108020111006	1	1	3,1	26	9	0,69	1,8	0,74	2,9	1,381
010802000301239338-1	1	1	3,4	29	10	1,30	2,2	0,70	3,7	2,922
010802000301239337-1	0	1	3,3	32	11	1,10	2,1	0,67	3,7	3,363
010802000301234958-1	0	1	3,1	35	12	1,30	2,1	0,66	4,0	3,872
010802000301234959-1	0	1	2,9	38	13	0,71	1,9	0,63	2,8	1,985
0108020110139-1	1	1	3,1	41	14	0,26	2,0	0,60	2,7	1,274
010802000301234951-1	1	1	3,2	43	15	0,43	2,1	0,57	2,7	2,058
010802000301234848-1	1	1	3,3	45	16	0,89	2,3	0,55	2,6	2,066
0108020111019	1	1	3,4	49	17	0,17	2,4	0,54	2,6	1,370

Tabel A.6: Testforløbdata for elev 129172, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020111018	0	1	1,500	2	1	0,553	-0,247	2,194	2,069	1,017
010802000301234952-1	0	1	1,000	5	2	0,206	-0,539	1,897	1,497	1,073
0108020111011	0	1	-1,121	8	3	0,204	-1,430	1,819	0,945	-0,134
010802000301234955-1	0	1	-2,711	11	4	0,122	-2,350	1,802	-1,138	-0,994
0108020115060	0	1	-4,371	14	5	0,099	-2,700	1,704	-2,804	-0,719
010802000301237986-1	1	1	-3,841	17	6	0,117	-2,075	1,128	-4,503	-2,076
0108020110177-1	1	1	-3,218	20	7	0,148	-1,561	0,917	-3,964	-1,545
010802000301237983-1	1	1	-2,707	23	8	0,089	-1,424	0,844	-3,190	-2,446
0108020110172-1	1	1	-2,271	26	9	0,247	-1,143	0,770	-2,635	-1,386
0108020110228	1	1	-1,894	29	10	0,113	-0,763	0,726	-2,216	-0,223
0108020110111-1	1	1	-1,555	32	11	0,088	-0,605	0,689	-1,828	-1,152
0108020110052-1	1	1	-1,254	35	12	0,106	-0,427	0,658	-1,545	-0,795
0108020110007-1	1	1	-0,967	37	13	0,154	-0,283	0,634	-1,183	-0,766
0108020111012	1	1	-0,702	39	14	0,164	-0,100	0,614	-0,911	-0,336
0108020110053-1	1	1	-0,449	41	15	0,115	0,007	0,598	-0,606	-0,901
010802000301234962-1	1	1	-0,219	43	16	0,163	0,105	0,584	-0,422	-0,757
0108020110261	1	1	0,004	44	17	0,179	0,251	0,573	-0,142	0,046
0108020110245	1	1	0,212	45	18	0,132	0,330	0,562	0,050	-0,808
0108020110038-1	0	1	0,032	46	19	0,095	0,169	0,529	0,269	0,067
0108020110219-1	1	1	0,209	47	20	0,151	0,290	0,520	0,126	0,137
0108020110155-1	1	1	0,376	48	21	0,152	0,386	0,512	0,266	-0,175
0108020110232	0	1	0,230	49	22	0,108	0,219	0,486	0,434	-0,518
0108020115013	1	1	0,373	52	23	0,334	0,320	0,479	0,285	0,005
0108020110082-1	0	1	0,243	55	24	0,153	0,208	0,459	0,384	0,207
0108020110059-1	1	1	0,370	58	25	0,180	0,319	0,454	0,307	0,465
0108020110003-1	1	1	0,493	61	26	0,104	0,373	0,449	0,459	-0,630
010802000301234881-1	1	1	0,609	64	27	0,114	0,463	0,444	0,539	0,274
0108020111034	1	1	0,722	67	28	0,181	0,536	0,439	0,652	0,058
0108020110217-1	0	1	0,620	70	29	0,195	0,435	0,422	0,787	0,204
010802000301239438-1	1	1	0,724	73	30	0,192	0,523	0,419	0,704	0,524
010802000301234811-1	0	1	0,633	76	31	0,281	0,443	0,405	0,805	0,500
0108020110024-1	0	1	0,548	79	32	0,133	0,371	0,393	0,732	0,513
0108020110105-1	1	1	0,636	82	33	0,159	0,402	0,390	0,592	-0,887
0108020110162-1	0	1	0,560	85	34	0,434	0,357	0,382	0,789	1,167
0108020110234	0	1	0,483	88	35	0,088	0,266	0,372	0,596	-0,301
0108020110205-1	1	1	0,575	91	36	0,218	0,325	0,368	0,936	0,015
0108020110079-1	1	1	0,661	94	37	0,134	0,386	0,364	0,916	0,221
0108020110018-1	1	1	0,739	97	38	0,116	0,438	0,361	0,847	0,002
0108020111031	0	1	0,675	100	39	0,148	0,365	0,352	0,878	0,059
0108020110156-1	0	1	0,615	103	40	0,279	0,293	0,345	0,864	-0,089
0108020115028	0	1	0,566	106	41	0,188	0,235	0,338	1,113	0,193
0108020110085-1	1	1	0,642	109	42	0,141	0,284	0,335	1,009	0,059
010802000301234855-1	0	1	0,592	112	43	0,130	0,232	0,329	0,990	0,278

Tabel A.6: Testforløbdata for elev 129172, profilområde 2 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020110206-1	1	1	0,663	115	44	0,215	0,288	0,326	1,032	0,357
0108020110144-1	0	1	0,614	118	45	0,177	0,240	0,321	0,916	0,361
0108020110137-1	1	1	0,681	121	46	0,139	0,303	0,318	1,035	0,815
0108020110131-1	1	1	0,740	124	47	0,137	0,350	0,316	0,877	0,244
0108020110006-1	1	1	0,798	127	48	0,125	0,381	0,313	0,904	-0,490
0108020110088-1	1	1	0,853	130	49	0,335	0,429	0,311	0,892	0,496
0108020110036-1	0	1	0,803	133	50	0,210	0,380	0,306	0,904	0,322
0108020111023	1	1	0,855	136	51	0,117	0,421	0,303	0,899	0,216
0108020111033	1	1	0,906	139	52	0,144	0,459	0,301	0,945	0,115
0108020110074-1	0	1	0,860	142	53	0,105	0,416	0,296	0,979	0,396
0108020115027	1	1	0,911	145	54	0,097	0,455	0,294	1,057	0,272
0108020110164-1	1	1	0,960	148	55	0,126	0,474	0,293	1,045	-0,812
0108020110267	0	1	0,919	151	56	0,164	0,433	0,289	1,130	0,467
0108020003012349 56-1	0	1	0,884	154	57	0,193	0,393	0,285	1,305	0,415
0108020003012349 46-1	1	1	0,936	157	58	0,123	0,430	0,283	1,353	0,247
0108020111004	1	1	0,984	160	59	0,322	0,459	0,281	1,205	-0,059
0108020110263	0	1	0,946	163	60	0,139	0,425	0,277	1,155	0,619
0108020110201-1	1	1	0,993	166	61	0,150	0,460	0,275	1,273	0,287

Tabel A.7: Testforløbdata for elev 104649, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108030003012383 03-1	0	1	0,25	3	1	0,082	-1,217	2,360	0,585	-0,201
0108030003012388 55-1	0	1	0,00	6	2	0,093	-1,441	1,967	0,272	0,822
01080306061340014	0	1	-2,32	9	3	0,497	-1,910	1,783	-0,097	-0,255
0108030060613252-4	0	1	-3,94	12	4	0,223	-2,245	1,691	-2,377	-0,294
0108030320029	0	1	-5,85	15	5	0,423	-3,571	1,865	-4,356	-2,351
0108030606136 01-3	1	1	-4,24	18	6	1,153	-2,571	1,145	-3,409	-2,229
01080306061330109	1	1	-3,36	21	7	0,683	-1,953	0,929	-2,987	-1,567
0108030320040	1	1	-2,79	24	8	1,140	-1,461	0,817	-2,541	-0,915
0108030320022-3	1	1	-2,44	27	9	0,623	-1,201	0,747	-2,715	-1,406
01080306912-1_2	1	1	-2,13	30	10	0,752	-0,921	0,697	-2,330	-0,731
01080306061330104	1	1	-1,86	33	11	0,691	-0,751	0,661	-2,107	-1,228
0108030610700-3	3	4	-1,28	36	12	1,456	-0,293	0,586	-1,082	0,160
0108030311017	0	1	-1,45	39	13	1,315	-0,428	0,556	-1,232	-0,063
0108030610705-1	0	1	-1,60	42	14	0,104	-0,630	0,531	-1,366	-1,371
0108030320069-1	1	1	-1,46	43	15	1,276	-0,485	0,519	-1,558	-0,396
0108030320045	1	1	-1,32	46	16	2,141	-0,383	0,507	-1,444	-0,777
0108030610352	1	1	-1,19	49	17	0,399	-0,273	0,497	-1,312	-0,550
0108030320042	1	1	-1,06	52	18	0,925	-0,206	0,488	-1,164	-1,098
01080306061330069-2	1	3	-1,14	55	19	1,291	-0,363	0,434	-0,897	-0,500
01080306061330114	1	1	-1,04	58	20	0,608	-0,297	0,427	-1,112	-0,898

Tabel A.8: Testforløbdata for elev 349294, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108030003012413 90-1	0	1	0,25	3	1	2,50	-1,30	2,30	0,570	-0,275
01080306061330123	0	1	0,00	6	2	2,70	-1,80	1,90	0,330	-0,116
0108030606132 62-4	0	1	-2,20	9	3	2,50	-2,30	1,80	0,036	-0,570
01080306912-1_2	0	1	-3,90	12	4	1,80	-2,70	1,70	-2,300	-0,731
0108030606136 01-3	1	1	-3,10	15	5	0,96	-2,10	1,10	-3,400	-2,229
01080306061330109	1	1	-2,50	18	6	1,50	-1,60	0,92	-3,000	-1,567
0108030320040	1	1	-2,00	21	7	1,90	-1,20	0,81	-2,500	-0,915
0108030320030	1	1	-1,60	24	8	3,60	-1,00	0,75	-1,900	-1,309
0108030311006	1	1	-1,30	27	9	2,80	-0,76	0,70	-1,600	-0,954
0108030311017	1	1	-0,98	30	10	2,80	-0,49	0,67	-1,200	-0,063
01080306061340013-1	2	3	-0,78	33	11	6,80	-0,41	0,55	-0,950	-0,689

Tabel A.8: Testforløbdata for elev 349294, profilområde 3 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108030320041	1	1	-0,61	36	12	2,50	-0,31	0,54	-0,740	-0,814
010803000301238348-1	1	1	-0,44	39	13	1,60	-0,24	0,53	-0,550	-1,280
01080303060910323	1	1	-0,28	42	14	2,20	-0,13	0,52	-0,420	-0,404
010803000301241921-1	2	5	-0,31	45	15	45,00	-0,33	0,44	-0,019	-0,487

Tabel A.9: Testforløbdata for elev 259724, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020115072	0	1	1,50	2	1	0,43	0,032	2,20	1,90	1,377
0108020110081-1	0	1	1,00	5	2	0,57	-0,850	2,00	1,50	0,353
0108020111011	0	1	-1,10	8	3	2,90	-1,600	1,80	0,94	-0,134
0108020110086-1	0	1	-2,70	11	4	0,37	-2,200	1,70	-1,10	-0,597
0108020110172-1	1	1	-2,10	14	5	0,46	-1,400	1,10	-2,60	-1,386
0108020115042	1	1	-1,40	17	6	2,90	-0,980	0,92	-2,00	-1,020
0108020110147-1	1	1	-0,96	20	7	1,50	-0,550	0,82	-1,40	-0,230
0108020111012	1	1	-0,55	23	8	1,70	-0,250	0,76	-0,91	-0,336
0108020110190-1	1	1	-0,20	26	9	1,40	-0,096	0,72	-0,47	-0,876
0108020110084-1	1	1	0,12	29	10	2,60	0,078	0,68	-0,10	-0,527

Tabel A.10: Testforløbdata for elev 143590, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108030320047-2	0	1	0,250	3	1	0,722	-1,146	2,134	0,544	0,272
01080306061330121	0	1	0,000	6	2	2,562	-1,454	1,901	0,261	0,057
01080306061330122	0	1	-2,253	9	3	1,655	-1,983	1,760	0,090	-0,263
01080306912-1_2	1	1	-1,727	12	4	0,894	-1,088	1,113	-2,330	-0,731
0108030310024-3	1	1	-1,129	15	5	1,065	-0,833	0,954	-1,728	-1,541
0108030610396	1	1	-0,696	18	6	0,552	-0,573	0,857	-1,150	-1,118
01080303060910327	3	4	-0,262	21	7	2,380	-0,365	0,601	-0,549	-0,718
010803000301241593-1	5	5	0,466	24	8	9,072	0,239	0,576	-0,065	-0,466
01080306061340005-2	1	1	0,618	27	9	1,252	0,343	0,570	0,476	-0,493
01080303060940009-1	2	3	0,700	30	10	1,999	0,543	0,543	0,235	0,425

Tabel A.11: Testforløbdata for elev 317854, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020111018	1	1	2,5	2	1	0,18	1,8	2,50	2,1	1,017
010802000301234810-1	1	1	3,0	5	2	0,56	2,9	1,90	2,6	1,433
0108020111006	1	1	5,1	8	3	0,84	3,3	1,80	2,9	1,381
010802000301239468-1	0	1	4,5	11	4	0,47	3,1	1,30	5,0	4,022
0108020110231	0	1	3,9	14	5	0,65	2,5	1,00	4,5	2,768
010802000301234818-1	0	1	3,5	17	6	0,39	2,3	0,93	4,1	3,185
0108020110166-1	0	1	3,2	20	7	0,72	1,9	0,84	3,5	2,089
0108020110268	0	1	2,9	23	8	0,23	1,7	0,79	3,3	2,251
010802000301238021-1	1	1	3,2	26	9	0,25	1,9	0,73	3,0	1,323
010802000301239338-1	1	1	3,5	29	10	0,44	2,2	0,69	3,7	2,922
010802000301239337-1	0	1	3,3	32	11	0,65	2,1	0,67	3,7	3,363
010802000301234958-1	0	1	3,2	35	12	0,32	2,1	0,65	4,0	3,872
010802000301234959-1	0	1	2,9	38	13	0,49	1,9	0,62	2,8	1,985
0108020110139-1	1	1	3,1	41	14	0,34	2,0	0,60	2,7	1,274
010802000301234951-1	1	1	3,2	44	15	0,93	2,2	0,57	2,7	2,058
010802000301234848-1	1	1	3,3	46	16	0,30	2,3	0,55	2,6	2,066
0108020111019	1	1	3,4	49	17	0,37	2,4	0,53	2,6	1,370
0108020111022	0	1	3,2	52	18	0,49	2,2	0,51	2,5	1,532
0108020111026	1	1	3,3	55	19	0,46	2,2	0,50	2,3	0,837
010802000301234814-1	0	1	3,1	58	20	0,28	2,1	0,49	2,3	2,154

Tabel A.11: Testforløbdata for elev 317854, profilområde 2 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010802000301234953-1	1	1	3,2	61	21	0,50	2,2	0,48	2,2	1,411
0108020110266	1	1	3,2	64	22	0,37	2,2	0,47	2,2	0,544
010802000301234788-1	1	1	3,3	67	23	0,60	2,3	0,46	2,2	1,338
010802000301234941-1	0	1	3,1	70	24	0,39	2,1	0,45	2,2	1,359
0108020110150-1	1	1	3,2	73	25	0,54	2,2	0,44	2,1	1,108
0108020110083-1	0	1	3,1	76	26	1,40	2,1	0,43	2,1	1,275
0108020111016	1	1	3,1	79	27	0,28	2,1	0,42	2,0	0,452
010802000301234966-1	0	1	3,0	82	28	0,44	2,0	0,41	2,0	1,346
0108020115072	0	1	2,9	85	29	0,39	1,9	0,40	1,9	1,377
0108020111032	0	1	2,8	88	30	0,67	1,8	0,40	1,9	0,266
0108020110027-1	0	1	2,6	91	31	0,49	1,7	0,39	1,8	1,486
0108020111024	0	1	2,5	94	32	0,39	1,6	0,38	1,8	0,382

Tabel A.12: Testforløbdata for elev 219768, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020111018	1	1	2,5	2	1	0,30	1,8	2,50	2,1	1,02
010802000301234810-1	1	1	3,0	5	2	0,87	2,9	1,90	2,6	1,43
0108020111006	1	1	5,1	8	3	0,92	3,3	1,80	2,9	1,38
010802000301239468-1	1	1	6,6	11	4	0,53	5,2	2,00	5,0	4,02
010802000301234973-1	1	1	7,0	14	5	1,10	5,7	1,80	6,1	3,87
010802000301234885-1	0	1	7,0	17	6	1,60	4,6	1,10	8,0	3,90
010802000301234876-1	0	1	6,0	20	7	0,72	3,9	0,93	4,7	2,96
010802000301239469-1	0	1	5,3	23	8	0,61	3,5	0,83	4,7	3,62
0108020110231	0	1	4,8	26	9	1,30	3,1	0,76	4,5	2,77
010802000301234818-1	0	1	4,4	29	10	1,30	2,9	0,72	4,1	3,19
010802000301234958-1	0	1	4,1	32	11	0,32	2,8	0,69	4,0	3,87
010802000301239338-1	0	1	3,8	35	12	0,44	2,6	0,66	3,7	2,92
010802000301239337-1	0	1	3,6	38	13	0,78	2,5	0,64	3,7	3,36
0108020110166-1	0	1	3,4	41	14	0,72	2,3	0,62	3,5	2,09
0108020110268	0	1	3,2	44	15	0,31	2,1	0,60	3,3	2,25

Tabel A.13: Testforløbdata for elev 305503, profilområde 3

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
010803060613300 75	1	1	0,75	3	1	1,7	1,00	2,50	0,51	0,269
01080306061330078	1	1	1,00	6	2	4,1	1,70	1,90	0,73	-0,015
010803000301238868-1	1	1	3,30	9	3	4,6	2,80	1,80	0,93	1,482
010803000301235585-2	0	1	2,80	12	4	3,0	2,30	1,30	3,40	2,855
010803000301235768-1	0	1	2,20	15	5	9,7	1,70	1,00	2,80	1,833
010803000301235607-2	0	1	1,70	18	6	5,6	1,20	0,90	2,00	1,100
010803000301238349-1	0	1	1,30	21	7	3,0	1,00	0,83	1,60	1,721
010803000301238838-1	0	1	1,00	24	8	4,4	0,88	0,79	1,40	2,135
01080306910-1__2	0	1	0,76	27	9	2,7	0,68	0,75	1,10	1,224
010803000301236033-2	0	1	0,52	30	10	4,0	0,40	0,73	0,86	0,398
010803000301237450-3	0	1	0,29	33	11	4,1	0,19	0,71	0,60	0,645

Tabel A.14: Testforløbdata for elev 386356, profilområde 1

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108010420130	1	1	1,000	1	1		2,500	2,10	0,460	1,107
010801000301238556-1	1	1	1,500	4	2	0,42	2,800	1,90	1,100	1,346
010801000301234839-1	1	1	3,600	7	3	0,57	3,300	1,80	1,500	1,591
0108010410315	0	1	2,800	10	4	0,63	2,000	1,10	3,000	0,201
010801000301239196-1	0	1	2,200	13	5	0,31	1,700	0,96	2,400	2,383

Tabel A.14: Testforløbdata for elev 386356, profilområde 1 *(continued)*

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108010000301234841-1	0	1	1,800	16	6	0,32	1,300	0,86	2,200	0,865
0108010410094	0	1	1,500	19	7	0,36	0,930	0,79	1,800	0,859
0108010420122	0	1	1,200	22	8	0,45	0,700	0,75	1,500	1,157
0108010420160	0	1	0,910	25	9	0,57	0,580	0,72	1,200	1,684
0108010000301238872-1	0	1	0,670	28	10	0,19	0,410	0,70	1,000	1,072
0108010415120	0	1	0,440	31	11	0,19	0,310	0,68	0,700	1,580
0108010000301238996-1	0	1	0,210	34	12	0,35	0,098	0,67	0,450	0,295
0108010000301238353-1	0	1	-0,015	37	13	0,66	-0,120	0,67	0,230	-0,070
0108010420142	0	1	-0,230	40	14	0,38	-0,240	0,65	0,068	0,688
0108010420150	0	1	-0,430	43	15	0,38	-0,430	0,65	-0,130	-0,073
0108010415129	0	1	-0,640	45	16	0,52	-0,670	0,65	-0,370	-0,862
0108010410230009	0	1	-0,850	47	17	0,31	-0,790	0,65	-0,570	0,080
0108010000301238279-1	1	1	-0,650	48	18	0,51	-0,650	0,60	-0,780	-1,076
0108010420027	0	1	-0,820	49	19	0,42	-0,830	0,59	-0,620	-0,731
0108010440028	1	1	-0,660	50	20	0,28	-0,690	0,55	-0,810	-0,924
0108010415182	0	1	-0,800	51	21	0,20	-0,860	0,55	-0,630	-0,964
0108010410230021	1	1	-0,660	52	22	0,36	-0,710	0,52	-0,730	-0,590

Tabel A.15: Testforløbdata for elev 439773, profilområde 2

Opgavenummer	Rigtige	Delsp.	Theta	Spnr.	Profilspnr.	Tid	theta-2017	SEM-2017	beta-DNT	beta-2017
0108020111016	1	1	2,5	2	1	1,30	1,9	2,10	2,0	0,452
0108020111022	1	1	3,0	5	2	1,70	2,8	2,00	2,5	1,532
0108020111006	1	1	5,1	8	3	0,59	3,2	1,80	2,9	1,381
0108020000301239468-1	1	1	6,6	11	4	1,30	5,2	2,00	5,0	4,022
0108020000301234973-1	0	1	5,8	14	5	1,70	4,1	1,20	6,1	3,868
0108020000301234876-1	0	1	4,9	17	6	0,76	3,3	1,00	4,7	2,956
0108020000301239469-1	0	1	4,4	20	7	1,00	3,0	0,90	4,7	3,618
0108020110231	0	1	4,1	23	8	2,20	2,7	0,83	4,5	2,768
0108020000301234818-1	0	1	3,8	26	9	0,86	2,5	0,77	4,1	3,185
0108020000301234958-1	0	1	3,6	29	10	0,95	2,4	0,75	4,0	3,872
0108020000301239337-1	0	1	3,3	32	11	0,55	2,2	0,72	3,7	3,363
0108020110166-1	0	1	3,1	35	12	0,74	2,0	0,69	3,5	2,089
0108020110268	0	1	2,9	38	13	0,38	1,8	0,67	3,3	2,251
0108020000301238021-1	0	1	2,7	41	14	0,61	1,6	0,66	3,0	1,323
0108020000301234959-1	0	1	2,5	43	15	1,20	1,4	0,64	2,8	1,985
0108020111019	0	1	2,4	45	16	0,60	1,2	0,63	2,6	1,370
0108020000301234951-1	0	1	2,2	47	17	0,26	1,1	0,61	2,7	2,058
0108020000301234814-1	0	1	2,0	49	18	0,68	1,0	0,60	2,3	2,154
0108020111018	1	1	2,2	51	19	0,18	1,2	0,56	2,1	1,017
0108020000301234953-1	1	1	2,3	53	20	0,46	1,3	0,53	2,2	1,411
0108020000301234810-1	1	1	2,5	57	21	1,00	1,5	0,51	2,6	1,433
0108020000301234848-1	1	1	2,6	60	22	0,34	1,6	0,48	2,6	2,066
0108020110139-1	1	1	2,7	63	23	0,18	1,7	0,47	2,7	1,274

B

Analyse af person-fit baseret på 2017-sværhedsgrader

B.1 Indledning

Dette appendiks gengiver analyser af en række enkeltstående testforløb. Bortset fra et af de test der benyttes til at afprøve tilpasningen mellem testforløbene og Rasch-modellen, har de metoder der benyttes til analyserne, været en del af teorien for Rasch-modeller siden starten at 70'erne. På den anden side er de metoder vi bruger til afprøvningen af tilpasningen mellem enkeltstående testforløb og Rasch-modellen, så vidt vi ved kun implementeret i det program, DIGRAM, som er benyttet til analyserne i dette bilag, og endda kun delvist beskrevet i brugervejledningen til programmet. Af den grund indleder vi dette appendiks med et afsnit der beskriver metoderne, således at andre der måtte have brug for det, selv kan implementere metoderne og kontrollere resultaterne.

I det efter følgende antages det at $A = \{X_1, , X_k\}$ er en mængde af items fra en Rasch-model. Items kan både være dikotome eller polytome. $x = (x_1, , x_k)$ er en vektor af item-responser. Sådanne vektorer omtales i det efterfølgende som responsmønstre.

Rasch-modellen antager at responserne er stokastisk uafhængige med sandsynligheder givet ved

$$Prob(X_i = x) = \frac{exp(x\theta - \sum_{j=1}^{x_i} \beta_{ij})}{\sum_{z=0}^{m_i} exp(z\theta - \sum_{j=1}^{z} \beta_{ij})} \tag{B.1}$$

I formel (B.1) er θ en såkaldt person-parameter der angiver hvor dygtig eleven er. β-parametrene er item-parametre. Item-parametrene omtales som regel som sværhedsgrader når der er tale om dikotome items, og som tærskel-værdier, når der er tale om polytome items. Tallet m_i er lig med den maksimale score som en elev kan opnå på det pågældende item. For dikotome items, hvor der kun skelnes mellem rigtige og forkerte svar, er $m = 1$ og β_{i1} omtales som opgavens sværhedsgrad, men m kan være større hvis resultatet af opgaven vurderes på en gradueret skala fra nul til m.

Tærskelværdierne er defineret fra værdien $X_i = 1$ op til værdien $X_i = m$. Når der er tale om dikotome items hvor der kun er én item-parameter, nøjes man med at omtale sværhedsgraden som β_i uden henvisning til at det er en tærskelværdi der svarer til $X_i = 1$. Det er i forbindelse med det efterfølgende bekvemt at indføre en ekstra "tærskelværdi", β_{i0}, der altid antages at være lig med 0 således at $\sum_{j=1}^{x} \beta_{ij} = \sum_{j=0}^{x} \beta_{ij}$ og således at antallet af item-parametre for både dikotome og polytome items er lig med $m_i + 1$.

I Rasch-modellen er den samlede score, $R = \sum_{1}^{k} X_i$, statistisk sufficient for personparameteren fordi den betingede fordeling af svarmønstret givet den samlede score, $R = r$, ikke afhænger af den ukendte personparameter, θ.

For at undersøge om et bestemt svarmønster passer til Rasch-modellen, skal man undersøge om der er et eller andet overraskende og grænsende til usandsynligt ved svarmønstret i forhold til det Rasch-modellen ville

forvente. I den forbindelse er den statistiske sufficiens en særdeles bekvem egenskab fordi den betyder at vi kan undersøge om svarmønstret er mere eller mindre usandsynligt, ved at beregne den betingede sandsynlighed for svarmønstret givet den samlede score uden at inddrage estimater af den ukendte personparameter.

I forbindelse med afprøvningen af tilpasningen mellem Rasch-modellen og svarmønstrene for enkelte elever har vi beregnet tre forskellige typer af betingede sandsynligheder, der på hver sin måde kan fortælle om et svarmønster – set fra Rasch-modellens synsvinkel – er overraskende og dermed udtryk for signifikant evidens mod tilpasningen.

Den første er den betingede sandsynlighed for det samlede svarmønster givet den samlede score,

$$Prob(X_1 = x_1, X_2 = x_2, ..., X_m = x_m | R = r) = \frac{Prob(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)}{Prob(R = r)} \tag{B.2}$$

Idet svarmønstrets tilpasning til Rasch-modellen forkastes hvis denne sandsynlighed er overraskende lille.

Da antallet af mulige svarmønstre med en givet samlet score kan være meget stort, og da summen sandsynlighederne for samtlige svarmønstre altid er lig med 1, må man forvente at de fleste eller måske alle betingede sandsynligheder er relativt små. For at tage stilling til om det observerede svarmønster afviger signifikant fra det Rasch-modellen forventer, beregnes derfor en såkaldt p-værdi der er givet som summen af sandsynlighederne for alle de svarmønstre der har samme eller endnu mindre sandsynlighed for at forekomme som det observerede svarmønster.

Formel (B.1) parametricerer Rasch-modellen som en såkaldt *partial credit-model* (PCM). Denne parametricering er populær fordi den placerer Rasch-modellen som et medlem af familien af item-respons-modeller, men parametriceringen er særdeles ubekvem hvis man vil udlede formler der gør det mulig at beregne de betingede sandsynligheder. De efterfølgende formler viser hvorledes man kan reformulere modellen således at formlerne både bliver mere læsbare og lettere at skrive op.

I det første skridt indføres nye item-parametre defineret ved $\sigma_{ix} = -\sum_{j=1}^{x} \beta_{ij}$ hvorefter PCM-formlen (B.1) omskrives til

$$Prob(X_i = x) = \frac{exp(x\theta + \sigma_{ix})}{\sum_{z=0}^{m_i} exp(z\theta + \sigma_{iz})} \tag{B.3}$$

For yderligere at forenkle formlerne er det herefter bekvemt at sætte $\delta_{ix} = exp(\sigma_{ix})$ således at (B.3) reduceres til

$$Prob(X_i = x) = \frac{exp(x\theta)\delta_{ix}}{\sum_{z=0}^{m_i} exp(z\theta)\delta_{iz}} \tag{B.4}$$

Med denne parametricering er det relativt let at se, at fordelingen af svarmønstret er givet ved sandsynlighederne

$$Prob(X_1 = x_1, X_2 = x_2, ..., X_m = x_m) = \frac{exp(r\theta)\prod_{i=1}^{k} \delta_{ix_i}}{G(\theta)} \tag{B.5}$$

hvor $r = \sum_{i=1}^{k} x_{ix_i}$ og $G(\theta) = \sum_{x=(x_i,...,x_k)} (exp(r\theta)\prod_{i=1}^{k} \delta_{ix_i})$ samt at sandsynligheden for at den samlede score lig med r er lig med

$$Prob(R=r) = \sum_{(x_1,...x_k):r=\sum_i x_i} Prob(X_1=x_1, X_2=x_2, ..., X_m=x_m) = \frac{exp(r\theta)}{G(\theta)} \sum_{(x_1,...x_k):r=\sum_i x_i} \prod_i \delta_{ix} \tag{B.6}$$

Hvis man derefter sætter $\gamma_r = \sum_{(x_1,...x_k):r=\sum_i x_i} \prod_i \delta_{ix}$ får man den formel for fordelingen af den samlede score over et sæt af items fra en Rasch-model som Rasch (1960) benyttede da teorien for Rasch-modellen blev defineret

$$Prob(R=r) = \frac{exp(r\theta)\gamma_r}{G(\theta)} \tag{B.7}$$

Hvorefter den betingede sandsynlighed for et bestemt svarmønster, (x_1, x_k), er lig med

$$Prob(X_1=x_1, X_2=x_2, ..., X_m=x_m|R=r) = \frac{\prod_{i=1}^{k} \delta_{ix_i}}{\gamma_r} \tag{B.8}$$

Hvis der kun er tale om dikotome items, omtales γ_r som det symmetriske polynomium af orden r. Med polytome items er dette ikke helt korrekt, men teorien for Rasch-modeller omtaler alligevel værdierne som gamma-polynomier.

I det foregående er gamma-polynomierne beregnet for hele mængden, A, af items, men polynomierne kan naturligvis beregnes for alle delmængder af items. Hvis man vil udlede formler hvor der optræder gamma-polynomier for delmængde af items, er det nødvendigt at formulere det således at man kan se, hvilke delmængder der er tale om. Hvis $B \subset A$ er en delmængde af items omtales gamma-polynomiet over disse items som $\gamma_r(B)$.

Dette får vi brug for i forbindelse med de næste to sandsynligheder som bruges til at afprøve om der er tilpasning mellem et konkret svarmønster og Rasch-modellen.

Den betingede fordeling af et enkelt item givet den samlede score på r er givet ved

$$Prob(X_i=x_i|R=r) = \frac{\delta_{ix_i}\gamma_{r-x_i}(A\backslash X_i)}{\gamma_r} \tag{B.9}$$

Disse sandsynligheder beregnes rutinemæssigt for samtlige items og kan i specielle tilfælde hvor der svares forkert på meget lette items eller korrekt på meget vanskelige items, afsløre at tilpasningen mellem svarmønstret og Rasch-modellen er utilstrækkeligt. Hvis det drejer sig om adaptive test vil algoritmen have udvalgt opgaver hvor der er ca. 50 procents chance for et korrekt svar – hvis sværhedsgraderne er korrekte og den adaptive algoritme fungerer efter hensigten. I sådanne tilfælde kan der ikke være små sandsynligheder for enkelte items, og samtlige svarmønstre vil have næste samme sandsynligheder for at forekomme. Sandsynlighederne (B.8) og (B.9) definerer derfor test med relativt begrænset styrke med mindre der ligefrem er fejl i forudsætningerne for den adaptive algoritme. Af den grund har vi udvidet testene baseret på (B.8) og (B.9) med test baseret på den samlede score i starten af forløbet.

Lad derfor $A_i = \{X_1, .., X_i\}$ være de første i opgaver og lad R_i være den samlede score over disse items.

Den betingede fordeling af R_i givet R er givet ved følgende sandsynligheder,

$$Prob(R_i=r_i|R=r) = \frac{\gamma_{r_i}(A_i)\gamma_{r-r_i}(A\backslash A_i)}{\gamma_r} \tag{B.10}$$

For at teste om der er signifikant få eller signifikant mange korrekte svar i forbindelse med de første i opgaver, beregnes herefter to kumulerede sandsynligheder – to p-værdier – for udfald af R_i henholdsvis op til og op fra den observerede r_i værdi $P_i = \sum_{x \leq r_i} Prob(R_i = x|R = r)$ og $Q_i = \sum_{x \geq r_i} Prob(R_i = x|R = r)$ hvorefter tilpasningen til Rasch-modellen forkastes, hvis p_i eller q_i er mindre end 0,025.

B.2 Resultater

Tabel B1 opsummerer resultaterne af tests af tilpasningen mellem forløbene og Rasch-modellen inklusive beregninger af de betingede sandsynligheder af de samlede forløb (søjle 2 og 3), af de enkelte items (søjle 4) og af vurderinger af starten af forløbet (søjle 5).

Tabel B1. Tests af tilpasning mellem svarmønstre og Rasch-modellen for 15 elever. Søjle 2 angiver den betingede sandsynlighed (8) for det samlede svarmønster. P-værdien knyttet til denne sandsynlighed ses i søjle 3. Søjle 4 viser sandsynligheden (9) for det item med den mindste sandsynlighed for det observerede udfald, mens søjle 5 viser den mindste p_i- eller q_i-værdi der er observeret for eleven.

	Response vector		Item	Subscore
Person	Conditional probability	p-value	p	p
441985	0.00009287364475177000	0.254	0.2381	0.1567
421613	0.00001789260290787000	0.250	0.1419	0.0001
129172	0.00000000000000200000	0.725	0.1862	0.0050
104649	0.00000577202527575000	0.493	0.1048	0.0063
349294	0.00038242827130304000	0.979	0.2304	0.0299
259724	0.02372005293312173000	0.957	0.3073	0.0237
143590	0.00293881490003593000	0.838	0.2372	0.0268
428314	0.00140118668466098000	0.963	0.1711	0.0590
219768	0.00076454121996100000	0.359	0.1098	0.0008
387213	0.11073991341957229000	0.796	0.3222	0.1107
439773	0.00002241574544920000	0.611	0.0800	0.0137
305503	0.02131733610277121000	0.748	0.1963	0.0213
386356	0.00000524481047458000	0.071	0.0774	0.0005
341070	0.00175769063066167000	0.907	0.3395	0.1473
317854	0.00000001464366035000	0.397	0.1871	0.0025

Da der er tale om adaptive test, kan der som forventet ikke påvises signifikante afvigelser fra de samlede forløb og for de enkelte opgaver. I elleve ud af femten tilfælde afviger starten på testforløbene fra det man ville forvente, hvis Rasch-modellen var gældende fra start til slut.

De hyppigste afvigelser fra Rasch-modellen i starten af testforløbet ses i tilfælde, hvor der startes med påfaldende mange fejl, hvorefter testadfærden tilsyneladende svarer til det Rasch-modellen ville forvente. Som konsekvens af dette er der i disse tilfælde foretaget nye beregninger af dygtigheden, hvor den mislykkede start på forløbene er blevet fraregnet. Resultatet af disse beregninger kan ses i tabel B2.

Tabel B2. WML estimater af dygtigheden baseret på korrekte 2017 sværhedsgrader (søjle 4) med eksakte beregninger af usikkerheden (søjle 5). I de tilfælde, hvor dele af forløbet tilsyneladende er mislykket er dygtigheden beregnet ud fra den del af forløbet, hvor testadfærden passer til Rasch-modellen. Disse tal ses i søjle 6 og 7. Den sidste søjle viser korrelationen mellem det aktuelle estimat af dygtigheden og sværhedsgraden af den efterfølgende opgave.

Person	Score	max	All items WML	SEM	Item subset WML	SEM	Person-Item correlation
441985	9	17	1.88	0.51			0.59
421613	13	21	-0.63	0.44	0.93	0.79	0.64
129172	38	61	0.46	0.28	0.70	0.30	0.73
104649	15	25	-0.30	0.43	0.32	0.56	0.69
349294	13	21	-0.33	0.44	0.20	0.56	0.71
259724	6	10	0.08	0.69			0.67
143590	14	19	0.54	0.55	1.33	0.74	0.67
428314	18	25	0.79	0.51			0.86
219768	5	15	2.10	0.61	-0.11	0.68	0.71
387213	4	10	2.03	0.76			0.86
439773	9	23	1.70	0.47	1.17	0.56	0.78
305503	3	11	0.19	0.72	-1.62	0.71	0.79
386356	6	22	-0.71	0.52	-1.57	0.66	0.80
341070	9	17	2.36	0.54			0.65
317854	15	32	1.56	0.38			0.64

Alle sandsynlighederne i tabel B1 og B2 er baseret på estimaterne af sværhedsgrader i 2017. Den efterfølgende tabel B3 viser resultater hvis DNT's item-parametre havde været benyttet. Bemærk at usikkerheden på estimaterne i tabel B3 ikke er påfaldende bedre end i tabel 2, men at korrelationen mellem de person-parametre, der estimeres undervejs, og sværhedsgraderne af de efterfølgende opgaver er væsentlig højere end i tabel 2. Den adaptive algoritme fungerer som den skal, men gevinsten er tilsyneladende begrænset.

Tabel B3. WML estimater af dygtigheden baseret på de sværhedsgrader som DNT benytter.

Person	Score	max	All items WML	SEM	Item subset WML	SEM	Person-Item correlation
441985	9	17	2.88	0.50			0.82
421613	13	21	-1.60	0.46	-0.16	0.78	0.82
129172	38	61	0.99	0.28	1.25	0.30	0.90
104649	15	25	-1.05	0.44	-0.50	0.54	0.94
349294	13	21	-0.32	0.45	0.09	0.55	0.98
259724	6	10	0.07	0.78			0.90
143590	14	19	0.64	0.52	1.27	0.70	1.00
428314	18	25	0.85	0.46	1.35	0.57	1.00
219768	5	15	3.23	0.61	0.95	0.68	0.92
387213	4	10	2.38	0.79			1.00
439773	9	23	2.74	0.47	2.16	0.56	0.97
305503	3	11	0.37	0.71	-1.46	0.70	1.00
386356	6	22	-0.62	0.52	-1.44	0.66	0.99
341070	9	17	3.38	0.52			0.73
317854	15	32	2.55	0.38			0.76

De efterfølgende sider indeholder forholdsvis uredigeret output fra analyserne, hvor man kan finde alle detaljer, inklusive de person-parametre der estimeres undervejs i forløbet, og sværhedsgrader for de opgaver som den adaptive rutine derefter vælger. For at hjælpe læseren er outputtet kommenteret for de første to elever, hvorefter vi håber at læseren selv kan gennemskue indholdet for de øvrige elever.

Den første tabel viser hvilke opgaver eleven har besvaret, om svaret var rigtigt (1) eller forkert (0) og hvor mange rigtige svar, der efterhånden samles op. Den sidste søjle med de såkaldte PCM thresholds angiver opgavernes sværhedsgrader. Eleven besvarer 9 ud af 17 dikotome opgaver korrekt.

Det er dette resultat, der benyttes til at beregne dygtigheden målt på Rasch-modellens logit-skala.

```
  +-----------------+
  |                 |
  | Elev nr. 441985 |
  |                 |
  +-----------------+

Profil = 2

Score = 9 out of 17
                                PCM
Item                    Score  Thresholds
0108020115072            1  1    1.38
010802000301234810-1     0  1    1.47
010802000301234966-1     1  2    1.37
010802000301234959-1     0  2    1.99
010802000301234953-1     1  3    1.42
010802000301234951-1     0  3    2.05
0108020111026            1  4    0.87
0108020110139-1          0  4    1.29
0108020111022            1  5    1.53
010802000301234848-1     0  5    2.09
0108020111019            1  6    1.37
010802000301238021-1     1  7    1.34
0108020111006            0  7    1.39
0108020110268            0  7    2.28
0108020110166-1          1  8    2.12
010802000301239338-1     1  9    2.98
010802000301239337-1     0  9    3.39
```

Derefter vises den betingede sandsynlighed for det observerede svarmønster givet en samlet score på 9 ud af 17 mulige point. Sandsynligheden er meget lille hvilket ikke er overraskende fordi man kan få 9 ud af 17 rigtige på 24.310 måder.

Den samlede sandsynlighed for alle svarmønstre med mindre eller samme sandsynlighed er lig med 0,254. Der er altså ikke tale om signifikant afvigelse fra Rasch-modellens forventninger.

```
Prob(1,0,1,0,1,0,1,0,1,0,1,1,0,0,1,1,0 | Score = 9) =  0.00009287
Monte Carlo test of person fit: p =   0.254
```

Den næste tabel viser hvad der sker undervejs. Dygtigheden (WML) og usikkerheden beregnes fra og med tredje opgave og vises sammen med sværhedsgraden for næste opgave. Efter sjette trin er dygtigheden lig med 1,16, SEM = 0,83 og sværhedsgraden af næste opgave er lig med 0,87. DNT vælger altså en ny opgave som ifølge sværhedsgraderne fra 2017 er noget lettere end den burde være.

```
                          Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 0108020115072         1  0.62974     1   0.62974
 2 010802000301234810-1  0  0.39136     1   0.50255
 3 010802000301234966-1  1  0.63323     2   0.48135         1.92  1.05     1.99
 4 010802000301234959-1  0  0.52844     2   0.40503         1.55  0.99     1.42
 5 010802000301234953-1  1  0.62098     3   0.41637         1.87  0.91     2.05
 6 010802000301234951-1  0  0.54597     3   0.36464         1.61  0.83     0.87
 7 0108020111026         1  0.74592     4   0.38689         1.77  0.78     1.29
 8 0108020110139-1       0  0.34737     4   0.28060         1.48  0.73     1.53
```

```
 9 0108020111022        1  0.59302     5   0.35803         1.69  0.69     2.09
10 010802000301234848-1 0  0.55414     5   0.27526         1.55  0.65     1.37
11 0108020111019        1  0.63404     6   0.35918         1.70  0.62     1.34
12 010802000301238021-1 1  0.64127     7   0.42594         1.83  0.60     1.39
13 0108020111006        0  0.37053     7   0.29732         1.65  0.57     2.28
14 0108020110268        0  0.60364     7   0.15666         1.56  0.55     2.12
15 0108020110166-1      1  0.43655     8   0.33911         1.73  0.53     2.98
16 010802000301239338-1 1  0.23805     9   0.83135         1.93  0.52     3.39
17 010802000301239337-1 0  0.83135     9   1.00000         1.88  0.51

Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item =  0.59

Estimates of person parameter

Optimal SEM with 17 dichotomous items ~  0.485

   WML =  1.883   SEM =  0.509   Bias =  0.001
    ML =  1.892   SEM =  0.539   Bias =  0.001

================================================
```

Korrelationen mellem WML-værdierne og næste sværhedsgrad er lig med 0,59.

Efter 17 opgaver vurderes dygtigheden på logit skalaen til at være 1,883 med en SEM på 0,509.

Bemærk, at vi både beregner WML-estimatet og det maksimum likelihood (ML) estimat som DNT benytter. Ifølge teorien burde der ikke være store forskelle i et adaptivt test. Det bekræftes af resultaterne.

```
  +-----------------+
  |                 |
  | Elev nr. 421613 |
  |                 |
  +-----------------+

Profil = 3

Score = 13 out of 21
                                  PCM
Item                    Score  Thresholds
01080306061340005-2      0  0   -0.47
0108030311018            0  0    0.19
01080306061330044-3      0  0    0.20
010803060613252-4        0  0   -0.26
0108030320029            0  0   -2.37
010803060613601-3        0  0   -2.33
01080306061330109        0  0   -1.59
0108030320022-3          1  1   -1.44
0108030320040            1  2   -0.91
01080306912-1_2          1  3   -0.78
0108030612006-1          1  4   -1.46
0108030320033            1  5   -1.54
01080306061330104        1  6   -1.22
0108030612019-1          0  6   -0.35
```

```
01080306061330034-2     1  7   -1.07
0108030310373           4 11   -0.89 -1.08 -1.30 -0.93   Target = -1.10  Info at target =  2.06
0108030320069-2         1 12   -0.93
01080303060940011-1     1 13   -1.03

Prob(0,0,0,0,0,0,0,1,1,1,1,1,1,0,1,4,1,1 | Score = 13) =  0.00001789

Monte Carlo test of person fit: p =   0.250
```

Den tredje sidste opgave er en polytom opgave med fire spørgsmål.

Target angiver den værdi af dygtigheden hvor denne opgave er mest informativ, og som DNT bruger til at vælge næste opgave.

Da en ud af 18 opgaver er polytom, kan denne elev score op til 21 point. Slutresultatet er imidlertid kun 13 point fordi der svares forkert på de første syv opgaver.

De mange forkerte svar i starten af forløbet ser mistænkelige ud, men selv om den betingede sandsynlighed for det observerede givet at der scores 13 ud af 21 mulige point, er meget lille, er det dog langt fra signifikant i sig selv. Set isoleret ud fra fokus på detaljerne er der intet der strider mod Rasch-modellens forventninger.

Det er der til gengæld hvis vi ser på sandsynligheden for at score nul point i de første opgaver. Nedenstående tabel viser således at sandsynligheden for at score nul point efter de første fem opgaver er nede på 0,013697. Dette tal er udtryk for signifikant forskel på starten af forløbet og Rasch-modellens forventninger, men det stopper ikke der.

Sandsynligheden for at score nul point på de første syv opgaver er endnu mindre, nemlig nede på 0,00010 altså endnu mere signifikant. Derefter begynder de korrekte svar at dukke op, men bemærk at det stadig er svagt signifikant, hvis der kun scores 4 point på de første 11 opgaver.

```
                        Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 01080306061340005-2  0  0.54490     0   0.54490
 2 0108030311018        0  0.70523     0   0.37428
 3 01080306061330044-3  0  0.70799     0   0.25294        -2.03  0.79    -0.26
 4 010803060613252-4    0  0.59885     0   0.13697        -2.33  0.74    -2.37
 5 0108030320029        0  0.14187     0   0.01359        -3.59  0.86    -2.33
 6 010803060613601-3    0  0.14689     0   0.00103        -4.08  0.79    -1.59
 7 01080306061330109    0  0.27019     0   0.00010        -4.23  0.75    -1.44
 8 0108030320022-3      1  0.69782     1   0.00079        -3.00  0.90    -0.91
 9 0108030320040        1  0.56871     2   0.00457        -2.35  0.84    -0.78
10 01080306912-1_2      1  0.53656     3   0.01799        -1.89  0.75    -1.46
11 0108030612006-1      1  0.70224     4   0.03655        -1.63  0.68    -1.54
12 0108030320033        1  0.71896     5   0.06353        -1.43  0.63    -1.22
13 01080306061330104    1  0.64785     6   0.11350        -1.25  0.60    -0.35
14 0108030612019-1      0  0.57593     6   0.04892        -1.33  0.58    -1.07
15 01080306061330034-2  1  0.60973     7   0.10353        -1.17  0.55    -1.10
16 0108030310373        4  0.40537    11   0.33251        -0.79  0.45    -0.93
17 0108030320069-2      1  0.57491    12   0.59974        -0.71  0.45    -1.03
18 01080303060940011-1  1  0.59974    13   1.00000        -0.63  0.44

Responses to items 1 to 7 are significant. p = 0.0001
Person parameter will be estimated based on the subscore of items 8 - 18

Rank correlation between person estimate and next item =  0.64
```

```
Optimal SEM with 21 dichotomous items ~  0.471

   WML = -0.631   SEM =  0.441   Bias =  0.004
    ML = -0.590   SEM =  0.478   Bias =  0.004

Subscore from 8 to 18 = 13  Max = 14

   WML =  0.932    SEM =  0.788
```

Ovenstående resultater forkaster Rasch-modellen for de første syv opgaver og fortæller derfor at det samlede bud på dygtigheden, der er lig med WML = -0,631, formodentlig undervurderer læsefærdigheden.

For at finde ud af hvor meget beregner programmet derfor dygtigheden hvis man kun bruger opgaverne 8 1il 18. Det giver et helt andet resultat, nemlig WML = 0,932.

I dette eksempel er konklusionen at man ved at ignorere det der sker i starten af forløbet, får et fuldstændig misvisende mål for hvor godt eleven læser. Det sidste bud er formodentlig tættere på, men her er usikkerheden til gengæld så stor, at dette resultat heller ikke er anvendeligt. Dette er et eksempel på et helt mislykket testforløb med et resultat der burde ignoreres.

B.3 Ukommenterede forløb

```
  +-----------------+
  |                 |
  | Elev nr. 129172 |
  |                 |
  +-----------------+

Profil = 2

Score = 38 out of 61

                                PCM
Item                    Score  Thresholds
0108020111018            0  0    1.05
010802000301234952-1     0  0    1.09
0108020111011            0  0   -0.07
010802000301234955-1     0  0   -0.99
0108020115060            0  0   -0.74
010802000301237986-1     1  1   -2.25
0108020110177-1          1  2   -1.45
010802000301237983-1     1  3   -2.63
0108020110172-1          1  4   -1.38
0108020110228            1  5   -0.09
0108020110111-1          1  6   -1.31
0108020110052-1          1  7   -0.84
0108020110007-1          1  8   -0.89
0108020111012            1  9   -0.25
0108020110053-1          1 10   -0.84
010802000301234962-1     1 11   -0.79
0108020110261            1 12   -0.04
```

```
0108020110245           1 13   -0.74
0108020110038-1         0 13   -0.05
0108020110219-1         1 14    0.04
0108020110155-1         1 15   -0.19
0108020110232           0 15   -0.51
0108020115013           1 16    0.05
0108020110082-1         0 16    0.13
0108020110059-1         1 17    0.43
0108020110003-1         1 18   -0.63
0108020003012348 81-1   1 19    0.27
0108020111034           1 20    0.03
0108020110217-1         0 20    0.21
0108020003012394 38-1   1 21    0.51
0108020003012348 11-1   0 21    0.50
0108020110024-1         0 21    0.53
0108020110105-1         1 22   -0.96
0108020110162-1         0 22    1.15
0108020110234           0 22   -0.34
0108020110205-1         1 23    0.05
0108020110079-1         1 24    0.22
0108020110018-1         1 25    0.02
0108020111031           0 25    0.06
0108020110156-1         0 25   -0.12
0108020115028           0 25    0.22
0108020110085-1         1 26    0.05
0108020003012348 55-1   0 26    0.29
0108020110206-1         1 27    0.40
0108020110144-1         0 27    0.39
0108020110137-1         1 28    0.78
0108020110131-1         1 29    0.21
0108020110006-1         1 30   -0.39
0108020110088-1         1 31    0.43
0108020110036-1         0 31    0.29
0108020111023           1 32    0.18
0108020111033           1 33    0.15
0108020110074-1         0 33    0.45
0108020115027           1 34    0.27
0108020110164-1         1 35   -0.80
0108020110267           0 35    0.48
0108020003012349 56-1   0 35    0.41
0108020003012349 46-1   1 36    0.28
0108020111004           1 37   -0.05
0108020110263           0 37    0.64
0108020110201-1         1 38    0.28

Prob(0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,0,1,1,1,1,0,1,0,0,1,0,0,1,1,1,
0,0,0,1,0,1,0,1,1,1,1,0,1,1,0,1,1,0,0,1,1,0,1 | Score = 38) =  0.0000000000000020

Monte Carlo test of person fit: p =   0.725

                        Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 0108020111018        0  0.64713     0   0.64713
 2 0108020003012349 52-1 0  0.65448     0   0.41957
```

3	0108020111011	0	0.36870	0	0.14931	-1.43	0.82	-0.99
4	010802000301234955-1	0	0.18624	0	0.02543	-2.35	0.82	-0.74
5	0108020115060	0	0.22835	0	0.00505	-2.70	0.76	-2.25
6	010802000301237986-1	1	0.93923	1	0.00813	-2.07	1.01	-1.45
7	0108020110177-1	1	0.87451	2	0.01550	-1.56	0.91	-2.63
8	010802000301237983-1	1	0.95771	3	0.01859	-1.42	0.85	-1.38
9	0108020110172-1	1	0.86558	4	0.02911	-1.14	0.77	-0.09
10	0108020110228	1	0.63663	5	0.06347	-0.76	0.73	-1.31
11	0108020110111-1	1	0.85787	6	0.08101	-0.61	0.69	-0.84
12	0108020110052-1	1	0.78944	7	0.10811	-0.43	0.66	-0.89
13	0108020110007-1	1	0.79815	8	0.13447	-0.28	0.64	-0.25
14	0108020111012	1	0.67213	9	0.17575	-0.10	0.62	-0.84
15	0108020110053-1	1	0.78910	10	0.19877	0.01	0.60	-0.79
16	010802000301234962-1	1	0.78019	11	0.21886	0.11	0.59	-0.04
17	0108020110261	1	0.62394	12	0.24413	0.25	0.58	-0.74
18	0108020110245	1	0.77210	13	0.25096	0.33	0.57	-0.05
19	0108020110038-1	0	0.37393	13	0.20809	0.17	0.53	0.04
20	0108020110219-1	1	0.60431	14	0.23055	0.29	0.53	-0.19
21	0108020110155-1	1	0.65970	15	0.23872	0.39	0.52	-0.51
22	0108020110232	0	0.27070	15	0.20184	0.22	0.49	0.05
23	0108020115013	1	0.60399	16	0.22133	0.32	0.48	0.13
24	0108020110082-1	0	0.41693	16	0.18609	0.21	0.46	0.43
25	0108020110059-1	1	0.50685	17	0.21282	0.32	0.46	-0.63
26	0108020110003-1	1	0.75143	18	0.22000	0.37	0.45	0.27
27	010802000301234881-1	1	0.54831	19	0.22312	0.46	0.45	0.03
28	0108020111034	1	0.60689	20	0.21437	0.54	0.44	0.21
29	0108020110217-1	0	0.43795	20	0.21971	0.44	0.42	0.51
30	010802000301239438-1	1	0.48668	21	0.21302	0.52	0.42	0.50
31	010802000301234811-1	0	0.51045	21	0.21793	0.44	0.41	0.53
32	0108020110024-1	0	0.51833	21	0.20862	0.37	0.39	-0.96
33	0108020110105-1	1	0.80835	22	0.21366	0.40	0.39	1.15
34	0108020110162-1	0	0.66945	22	0.20368	0.36	0.38	-0.34
35	0108020110234	0	0.30660	22	0.16593	0.27	0.37	0.05
36	0108020110205-1	1	0.60320	23	0.19041	0.32	0.37	0.22
37	0108020110079-1	1	0.56156	24	0.21034	0.39	0.37	0.02
38	0108020110018-1	1	0.60967	25	0.21939	0.44	0.36	0.06
39	0108020111031	0	0.39925	25	0.20446	0.37	0.35	-0.12
40	0108020110156-1	0	0.35700	25	0.16729	0.29	0.35	0.22
41	0108020115028	0	0.43870	25	0.12560	0.23	0.34	0.05
42	0108020110085-1	1	0.60343	26	0.15654	0.28	0.34	0.29
43	010802000301234855-1	0	0.45685	26	0.11366	0.23	0.33	0.40
44	0108020110206-1	1	0.51512	27	0.15325	0.29	0.33	0.39
45	0108020110144-1	0	0.48265	27	0.10967	0.24	0.32	0.78
46	0108020110137-1	1	0.42017	28	0.16037	0.30	0.32	0.21
47	0108020110131-1	1	0.56261	29	0.20002	0.35	0.32	-0.39
48	0108020110006-1	1	0.70416	30	0.22519	0.38	0.31	0.43
49	0108020110088-1	1	0.50788	31	0.25629	0.43	0.31	0.29
50	0108020110036-1	0	0.45721	31	0.23198	0.38	0.31	0.18
51	0108020111023	1	0.56961	32	0.26886	0.42	0.30	0.15
52	0108020111033	1	0.57923	33	0.28947	0.46	0.30	0.45
53	0108020110074-1	0	0.49809	33	0.28662	0.42	0.30	0.27
54	0108020115027	1	0.54703	34	0.32097	0.46	0.29	-0.80
55	0108020110164-1	1	0.78296	35	0.32989	0.47	0.29	0.48
56	0108020110267	0	0.50438	35	0.34893	0.43	0.29	0.41

```
57 010802000301234956-1 0  0.48788    35   0.29554         0.39  0.28     0.28
58 010802000301234946-1 1  0.54494    36   0.41722         0.43  0.28    -0.05
59 0108020111004        1  0.62624    37   0.51381         0.46  0.28     0.64
60 0108020110263        0  0.54602    37   0.54636         0.43  0.28     0.28
61 0108020110201-1      1  0.54636    38   1.00000         0.46  0.28

Responses to items 1 to 5 are significant. p = 0.0050
Person parameter will be estimated based on the subscore of items 6 - 61

Rank correlation between person estimate and next item =  0.73

Estimates of person parameter

Optimal SEM with 61 dichotomous items ~  0.256

   WML =  0.460   SEM =  0.276   Bias =  0.000
    ML =  0.467   SEM =  0.281   Bias =  0.000


Subscore from 6 to 61 = 38  Max = 56

   WML =  0.701    SEM =  0.296

================================================

  +-----------------+
  |                 |
  | Elev nr. 104649 |
  |                 |
  +-----------------+

Profil = 3

Score = 15 out of 25

                                PCM
Item                    Score  Thresholds
010803000301238303-1     0  0   -0.21
010803000301238855-1     0  0    0.84
01080306061340014        0  0   -0.18
010803060613252-4        0  0   -0.26
0108030320029            0  0   -2.37
010803060613601-3        1  1   -2.33
01080306061330109        1  2   -1.59
0108030320040            1  3   -0.91
0108030320022-3          1  4   -1.44
01080306912-1_2          1  5   -0.78
01080306061330104        1  6   -1.22
0108030610700-3          3  9   -1.29 -1.19  0.52  2.58   Target = -0.96  Info at target =  0.86
0108030311017            0  9   -0.15
0108030610705-1          0  9   -1.41
0108030320069-1          1 10   -0.39
0108030320045            1 11   -0.81
0108030610352            1 12   -0.54
```

```
0108030320042           1 13   -1.11
01080306061330069-2     1 14   -0.84 -0.81  0.13   Target = -0.62  Info at target =  1.01
01080306061330114       1 15   -0.86

Prob(0,0,0,0,0,1,1,1,1,1,1,3,0,0,1,1,1,1,1,1 | Score = 15) =  0.000005772

Monte Carlo test of person fit: p =   0.493

                       Conditional probabilites       Person estimates
Item                   Item score  cumulated score         WML   SEM     Next
 1 010803000301238303-1 0  0.52451     0   0.52451
 2 010803000301238855-1 0  0.76627     0   0.39362
 3 01080306061340014    0  0.53113     0   0.19521        -1.91  0.80    -0.26
 4 010803060613252-4    0  0.50974     0   0.08720        -2.25  0.75    -2.37
 5 0108030320029        0  0.10480     0   0.00631        -3.57  0.87    -2.33
 6 010803060613601-3    1  0.89136     1   0.01451        -2.57  1.02    -1.59
 7 01080306061330109    1  0.79353     2   0.03388        -1.95  0.92    -0.91
 8 0108030320040        1  0.65385     3   0.07812        -1.46  0.82    -1.44
 9 0108030320022-3      1  0.76695     4   0.11633        -1.20  0.75    -0.78
10 01080306912-1_2      1  0.62406     5   0.18430        -0.92  0.70    -1.22
11 01080306061330104    1  0.72435     6   0.23328        -0.75  0.66    -0.96
12 0108030610700-3      3  0.20789     9   0.33014        -0.29  0.59    -0.15
13 0108030311017        0  0.53938     9   0.31867        -0.43  0.56    -1.41
14 0108030610705-1      0  0.23890     9   0.21796        -0.63  0.53    -0.39
15 0108030320069-1      1  0.52435    10   0.30022        -0.49  0.52    -0.81
16 0108030320045        1  0.62983    11   0.34537        -0.38  0.51    -0.54
17 0108030610352        1  0.56224    12   0.34979        -0.27  0.50    -1.11
18 0108030320042        1  0.70017    13   0.31855        -0.21  0.49    -0.62
19 01080306061330069-2  1  0.23888    14   0.64259        -0.36  0.44    -0.86
20 01080306061330114    1  0.64259    15   1.00000        -0.30  0.43

Responses to items 1 to 5 are significant. p = 0.0063
Person parameter will be estimated based on the subscore of items 6 - 20

Rank correlation between person estimate and next item =  0.69

Estimates of person parameter

Optimal SEM with 25 dichotomous items ~  0.447

   WML = -0.297   SEM =  0.430   Bias =  0.001
    ML = -0.275   SEM =  0.451   Bias =  0.001


Subscore from 6 to 20 = 15  Max = 20

   WML =  0.318    SEM =  0.557

================================================

  +-----------------+
  |                 |
  | Elev nr. 349294 |
  |                 |
```

```
  +-----------------+

Profil = 3

Score = 13 out of 21
                                PCM
Item                    Score  Thresholds
010803000301241390-1     0  0   -0.27
01080306061330123        0  0   -0.14
010803060613262-4        0  0   -0.68
01080306912-1_2          0  0   -0.78
010803060613601-3        1  1   -2.33
01080306061330109        1  2   -1.59
0108030320040            1  3   -0.91
0108030320030            1  4   -1.46
0108030311006            1  5   -0.86
0108030311017            1  6   -0.15
01080306061340013-1      2  8   -0.65 -0.75 -0.51   Target = -0.66  Info at target =  1.22
0108030320041            1  9   -0.82
010803000301238348-1     1 10   -1.24
01080303060910323        1 11   -0.44
010803000301241921-1     2 13   -0.72 -1.64 -1.03  0.66  0.47  Target = -1.07  Info at target =  1.58

Prob(0,0,0,0,1,1,1,1,1,1,1,2,1,1,1,2 | Score = 13) =  0.00038243

Monte Carlo test of person fit: p =   0.979

                         Conditional probabilites       Person estimates
Item                     Item score  cumulated score         WML   SEM     Next
 1 010803000301241390-1 0  0.51914     0   0.51914
 2 01080306061330123    0  0.55203     0   0.27422
 3 010803060613262-4    0  0.41096     0   0.09931        -2.34  0.78    -0.78
 4 01080306912-1_2      0  0.38545     0   0.02990        -2.70  0.74    -2.33
 5 010803060613601-3    1  0.88782     1   0.04595        -2.12  1.01    -1.59
 6 01080306061330109    1  0.78727     2   0.08041        -1.63  0.91    -0.91
 7 0108030320040        1  0.64476     3   0.14818        -1.22  0.81    -1.46
 8 0108030320030        1  0.76411     4   0.19642        -1.00  0.75    -0.86
 9 0108030311006        1  0.63317     5   0.26703        -0.76  0.71    -0.15
10 0108030311017        1  0.45019     6   0.33420        -0.49  0.68    -0.66
11 01080306061340013-1  2  0.33986     8   0.36182        -0.41  0.56    -0.82
12 0108030320041        1  0.62251     9   0.38531        -0.31  0.55    -1.24
13 010803000301238348-1 1  0.71969    10   0.37477        -0.24  0.54    -0.44
14 01080303060910323    1  0.52773    11   0.23041        -0.13  0.53    -1.07
15 010803000301241921-1 2  0.23041    13   1.00000        -0.33  0.44

Responses to items 1 to 4 are significant. p = 0.0299
Person parameter will be estimated based on the subscore of items 5 - 15

Rank correlation between person estimate and next item =  0.71

Estimates of person parameter

Optimal SEM with 21 dichotomous items ~  0.516
```

```
   WML = -0.334   SEM =  0.440   Bias =  0.002
    ML = -0.304   SEM =  0.470   Bias =  0.002

Subscore from 5 to 15 = 13  Max = 17

   WML =  0.200    SEM =  0.562

============================================

  +-----------------+
  |                 |
  | Elev nr. 259724 |
  |                 |
  +-----------------+

Profil = 2

Score = 6 out of 10

                              PCM
Item                    Score  Thresholds
0108020115072            0  0    1.38
0108020110081-1          0  0    0.40
0108020111011            0  0   -0.07
0108020110086-1          0  0   -0.65
0108020110172-1          1  1   -1.38
0108020115042            1  2   -1.04
0108020110147-1          1  3   -0.18
0108020111012            1  4   -0.25
0108020110190-1          1  5   -0.92
0108020110084-1          1  6   -0.45

Prob(0,0,0,0,1,1,1,1,1,1 | Score = 6) =  0.02372005

Monte Carlo test of person fit: p =   0.957

                        Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 0108020115072        0  0.81003     0   0.81003
 2 0108020110081-1      0  0.59122     0   0.46119
 3 0108020111011        0  0.45998     0   0.17119        -1.56  0.81    -0.65
 4 0108020110086-1      0  0.30730     0   0.02372        -2.20  0.78    -1.38
 5 0108020110172-1      1  0.83285     1   0.04711        -1.43  1.00    -1.04
 6 0108020115042        1  0.77712     2   0.09019        -0.98  0.92    -0.18
 7 0108020110147-1      1  0.57095     3   0.23137        -0.55  0.83    -0.25
 8 0108020111012        1  0.58946     4   0.46329        -0.25  0.76    -0.92
 9 0108020110190-1      1  0.75370     5   0.64445        -0.10  0.72    -0.45
10 0108020110084-1      1  0.64445     6   1.00000         0.08  0.69

Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item =  0.67

Estimates of person parameter
```

```
Optimal SEM with 10 dichotomous items ~  0.632

   WML =  0.078   SEM =  0.690   Bias =  0.002
    ML =  0.119   SEM =  0.776   Bias =  0.002

============================================

  +-----------------+
  |                 |
  | Elev nr. 143590 |
  |                 |
  +-----------------+

Profil = 3

Score = 14 out of 19
                                PCM
Item                    Score  Thresholds
0108030320047-2          0  0    0.27
01080306061330121        0  0    0.05
01080306061330122        0  0   -0.33
01080306912-1_2          1  1   -0.78
0108030310024-3          1  2   -1.55
0108030610396            1  3   -1.13
01080303060910327        3  6   -0.93 -0.95 -0.74 -0.02   Target = -0.78  Info at target =  1.63
010803000301241593-1     5 11   -0.97 -1.52 -1.15  0.61  0.71  Target = -1.14  Info at target =  1.54
01080306061340005-2      1 12   -0.47
01080303060940009-1      2 14   -1.17  0.84  1.82   Target =  1.02  Info at target =  0.64

Prob(0,0,0,1,1,1,3,5,1,2 | Score = 14) =  0.00293881

Monte Carlo test of person fit: p =   0.838

                         Conditional probabilites       Person estimates
Item                     Item score  cumulated score         WML   SEM     Next
 1 0108030320047-2       0  0.42541     0   0.42541
 2 01080306061330121     0  0.36953     0   0.13790
 3 01080306061330122     0  0.27983     0   0.02679        -1.98  0.78    -0.78
 4 01080306912-1_2       1  0.80605     1   0.05918        -1.09  0.98    -1.55
 5 0108030310024-3       1  0.90187     2   0.07915        -0.83  0.95    -1.13
 6 0108030610396         1  0.85607     3   0.11132        -0.57  0.86    -0.78
 7 01080303060910327     3  0.33444     6   0.04683        -0.36  0.62    -1.14
 8 010803000301241593-1  5  0.23718    11   0.28431         0.24  0.59    -0.47
 9 01080306061340005-2   1  0.74967    12   0.36339         0.34  0.58     1.02
10 01080303060940009-1   2  0.36339    14   1.00000         0.54  0.55

Responses to items 1 to 3 are significant. p = 0.0268
Person parameter will be estimated based on the subscore of items 4 - 10

Rank correlation between person estimate and next item =  0.67

Estimates of person parameter
```

```
Optimal SEM with 19 dichotomous items ~  0.632

   WML =  0.543   SEM =  0.550   Bias = -0.002
    ML =  0.601   SEM =  0.608   Bias = -0.002


Subscore from 4 to 10 = 14  Max = 16

   WML =  1.334    SEM =  0.745

================================================

  +-----------------+
  |                 |
  | Elev nr. 428314 |
  |                 |
  +-----------------+

Profil = 3

Score = 18 out of 25

                                PCM
Item                    Score  Thresholds
010803000301238841-1     0  0    0.89
0108030320023-2          0  0    0.57
01080306061330044-3      0  0    0.20
01080306912-1_2          1  1   -0.78
0108030310024-3          1  2   -1.55
01080306061330114        1  3   -0.86
0108030310611-2          5  8   -0.73 -0.87 -1.21 -0.99  0.01  Target = -0.95  Info at target =  2.63
010803000301239063-1     2 10   -1.96 -0.02 -0.34  1.95  2.15  Target = -0.13  Info at target =  1.03
010803000301241409-1     4 14   -0.58 -2.08 -0.18 -0.22  1.83  Target = -0.98  Info at target =  1.46
0108030310606-1          1 15   -0.54
010803000301237450-3     1 16    0.60
010803000301235608-1     1 17    0.60
010803000301238350-1     1 18    1.30

Prob(0,0,0,1,1,1,5,2,4,1,1,1,1 | Score = 18) =  0.00140119

Monte Carlo test of person fit: p =   0.963

                        Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 010803000301238841-1 0  0.52839     0   0.52839
 2 0108030320023-2      0  0.44298     0   0.21595
 3 01080306061330044-3  0  0.34925     0   0.05896        -1.44  0.78    -0.78
 4 01080306912-1_2      1  0.84025     1   0.09778        -0.73  1.02    -1.55
 5 0108030310024-3      1  0.92031     2   0.11822        -0.53  0.98    -0.86
 6 01080306061330114    1  0.85059     3   0.15776        -0.25  0.89    -0.95
 7 0108030310611-2      5  0.66541     8   0.25591         0.15  0.70    -0.13
 8 010803000301239063-1 2  0.17107    10   0.06713         0.07  0.56    -0.98
 9 010803000301241409-1 4  0.58918    14   0.06182         0.27  0.51    -0.54
10 0108030310606-1      1  0.80319    15   0.08604         0.34  0.51     0.60
```

```
11 010803000301237450-3 1  0.54894    16   0.18388         0.49  0.51     0.60
12 010803000301235608-1 1  0.55021    17   0.36424         0.62  0.51     1.30
13 010803000301238350-1 1  0.36424    18   1.00000         0.79  0.51

Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item =  0.86

Estimates of person parameter

Optimal SEM with 25 dichotomous items ~  0.555

   WML =  0.787   SEM =  0.506   Bias = -0.001
    ML =  0.833   SEM =  0.524   Bias = -0.001

================================================

  +-----------------+
  |                 |
  | Elev nr. 219768 |
  |                 |
  +-----------------+

Profil = 2

Score = 5 out of 15

                              PCM
Item                    Score  Thresholds
0108020111018             1  1    1.05
010802000301234810-1      1  2    1.47
0108020111006             1  3    1.39
010802000301239468-1      1  4    4.05
010802000301234973-1      1  5    3.93
010802000301234885-1      0  5    4.00
010802000301234876-1      0  5    2.96
010802000301239469-1      0  5    3.64
0108020110231             0  5    2.78
010802000301234818-1      0  5    3.19
010802000301234958-1      0  5    3.90
010802000301239338-1      0  5    2.98
010802000301239337-1      0  5    3.39
0108020110166-1           0  5    2.12
0108020110268             0  5    2.28

Prob(1,1,1,1,1,0,0,0,0,0,0,0,0,0,0 | Score = 5) =  0.00076454

Monte Carlo test of person fit: p =   0.359

                         Conditional probabilites       Person estimates
Item                    Item score  cumulated score         WML   SEM     Next
 1 0108020111018          1  0.75798     1   0.75798
 2 010802000301234810-1   1  0.66739     2   0.49076
 3 0108020111006          1  0.68704     3   0.31159         3.27  0.78     4.05
```

```
 4 010802000301239468-1 1  0.10976     4   0.02128         5.16  0.99     3.93
 5 010802000301234973-1 1  0.12237     5   0.00076         5.65  0.83     4.00
 6 010802000301234885-1 0  0.88546     5   0.00243         4.58  1.00     2.96
 7 010802000301234876-1 0  0.72038     5   0.01016         3.87  0.93     3.64
 8 010802000301239469-1 0  0.84127     5   0.01911         3.53  0.83     2.78
 9 0108020110231        0  0.67741     5   0.04900         3.14  0.76     3.19
10 010802000301234818-1 0  0.76640     5   0.08473         2.91  0.72     3.90
11 010802000301234958-1 0  0.87391     5   0.10948         2.80  0.69     2.98
12 010802000301239338-1 0  0.72368     5   0.18226         2.61  0.66     3.39
13 010802000301239337-1 0  0.80339     5   0.24981         2.49  0.64     2.12
14 0108020110166-1      0  0.50493     5   0.54772         2.27  0.62     2.28
15 0108020110268        0  0.54772     5   1.00000         2.10  0.61

Responses to items 1 to 5 are significant. p = 0.0008
Person parameter will be estimated based on the subscore of items 6 - 15

Rank correlation between person estimate and next item =  0.71

Estimates of person parameter

Optimal SEM with 15 dichotomous items ~  0.516

   WML =  2.099   SEM =  0.608   Bias = -0.000
    ML =  2.056   SEM =  0.658   Bias = -0.000

Subscore from 6 to 15 = 0  Max = 10

   WML = -0.114    SEM =  0.684

============================================

  +-----------------+
  |                 |
  | Elev nr. 387213 |
  |                 |
  +-----------------+

Profil = 3

Score = 4 out of 10

                                PCM
Item                    Score  Thresholds
010803000301238841-1     1  1    0.89
0108030320062-1          1  2    0.39
0108030320053-1          1  3    0.84
010803000301235582-2     1  4    2.67
010803000301235587-2     0  4    4.09
010803000301235606-2     0  4    3.25
010803000301235578-2     0  4    3.70
010803000301236038-1     0  4    3.33
010803000301235602-2     0  4    3.18
010803000301235608-2     0  4    2.33
```

```
Prob(1,1,1,1,0,0,0,0,0,0 | Score = 4) =  0.11073991

Monte Carlo test of person fit: p =   0.796
```

Item	Conditional probabilites: Item score	cumulated score		Person estimates: WML	SEM	Next
1 010803000301238841-1	1 0.78942	1	0.78942			
2 0108030320062-1	1 0.86542	2	0.67333			
3 0108030320053-1	1 0.79836	3	0.51600	2.68	0.78	2.67
4 010803000301235582-2	1 0.32219	4	0.11074	3.89	0.85	4.09
5 010803000301235587-2	0 0.90942	4	0.14884	3.48	1.16	3.25
6 010803000301235606-2	0 0.80380	4	0.24424	2.95	1.02	3.70
7 010803000301235578-2	0 0.86919	4	0.31638	2.72	0.92	3.33
8 010803000301236038-1	0 0.81696	4	0.43171	2.50	0.85	3.18
9 010803000301235602-2	0 0.79120	4	0.58481	2.30	0.80	2.33
10 010803000301235608-2	0 0.58481	4	1.00000	2.03	0.76	

```
Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item =  0.86

Estimates of person parameter

Optimal SEM with 10 dichotomous items ~  0.632

   WML =  2.026   SEM =  0.759   Bias = -0.002
    ML =  1.979   SEM =  0.815   Bias = -0.002

================================================

  +-----------------+
  |                 |
  | Elev nr. 439773 |
  |                 |
  +-----------------+

Profil = 2

Score = 9 out of 23
```

Item	Score	PCM Thresholds
0108020111016	1 1	0.46
0108020111022	1 2	1.53
0108020111006	1 3	1.39
010802000301239468-1	1 4	4.05
010802000301234973-1	0 4	3.93
010802000301234876-1	0 4	2.96
010802000301239469-1	0 4	3.64
0108020110231	0 4	2.78
010802000301234818-1	0 4	3.19
010802000301234958-1	0 4	3.90
010802000301239337-1	0 4	3.39
0108020110166-1	0 4	2.12

```
0108020110268           0  4    2.28
010802000301238021-1    0  4    1.34
010802000301234959-1    0  4    1.99
0108020111019           0  4    1.37
010802000301234951-1    0  4    2.05
010802000301234814-1    0  4    2.15
0108020111018           1  5    1.05
010802000301234953-1    1  6    1.42
010802000301234810-1    1  7    1.47
010802000301234848-1    1  8    2.09
0108020110139-1         1  9    1.29

Prob(1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1 | Score = 9) =  0.00002242

Monte Carlo test of person fit: p =   0.611
```

		Conditional probabilites			Person estimates		
Item		Item score	cumulated score		WML	SEM	Next
1	0108020111016	1 0.78583	1	0.78583			
2	0108020111022	1 0.54453	2	0.41834			
3	0108020111006	1 0.58200	3	0.22663	3.18	0.80	4.05
4	010802000301239468-1	1 0.08004	4	0.01371	5.16	1.00	3.93
5	010802000301234973-1	0 0.91062	4	0.03023	4.08	1.12	2.96
6	010802000301234876-1	0 0.79076	4	0.07271	3.32	1.00	3.64
7	010802000301239469-1	0 0.88350	4	0.09861	3.02	0.90	2.78
8	0108020110231	0 0.75660	4	0.15524	2.66	0.82	3.19
9	010802000301234818-1	0 0.82659	4	0.19538	2.46	0.77	3.90
10	010802000301234958-1	0 0.90786	4	0.21571	2.37	0.75	3.39
11	010802000301239337-1	0 0.85489	4	0.24648	2.25	0.72	2.12
12	0108020110166-1	0 0.60989	4	0.31902	2.00	0.70	2.28
13	0108020110268	0 0.64805	4	0.35660	1.82	0.68	1.34
14	010802000301238021-1	0 0.40575	4	0.34450	1.56	0.67	1.99
15	010802000301234959-1	0 0.57618	4	0.28145	1.41	0.65	1.37
16	0108020111019	0 0.41327	4	0.15926	1.22	0.64	2.05
17	010802000301234951-1	0 0.59312	4	0.08206	1.11	0.63	2.15
18	010802000301234814-1	0 0.61792	4	0.02884	1.02	0.62	1.05
19	0108020111018	1 0.66375	5	0.05406	1.18	0.57	1.42
20	010802000301234953-1	1 0.57321	6	0.11363	1.33	0.53	1.47
21	010802000301234810-1	1 0.56051	7	0.22822	1.46	0.51	2.09
22	010802000301234848-1	1 0.39903	8	0.60612	1.61	0.49	1.29
23	0108020110139-1	1 0.60612	9	1.00000	1.70	0.47	

```
Responses to items 1 to 4 are significant. p = 0.0137
Person parameter will be estimated based on the subscore of items 5 - 23

Rank correlation between person estimate and next item =  0.78

Estimates of person parameter

Optimal SEM with 23 dichotomous items ~  0.417

   WML =  1.698   SEM =  0.469   Bias =  0.000
    ML =  1.685   SEM =  0.489   Bias =  0.000
```

```
Subscore from 5 to 23 = 5  Max = 19

   WML =  1.167    SEM =  0.557

================================================

  +-----------------+
  |                 |
  | Elev nr. 305503 |
  |                 |
  +-----------------+

Profil = 3

Score = 3 out of 11

                                PCM
Item                    Score  Thresholds
01080306061330075         1  1    0.27
01080306061330078         1  2   -0.01
010803000301238868-1      1  3    1.45
010803000301235585-2      0  3    2.84
010803000301235768-1      0  3    1.84
010803000301235607-2      0  3    1.10
010803000301238349-1      0  3    1.72
010803000301238838-1      0  3    2.14
01080306910-1_2           0  3    1.26
010803000301236033-2      0  3    0.43
010803000301237450-3      0  3    0.60

Prob(1,1,1,0,0,0,0,0,0,0,0 | Score = 3) =  0.02131734

Monte Carlo test of person fit: p =   0.748

                          Conditional probabilites       Person estimates
Item                      Item score  cumulated score         WML   SEM     Next
 1 01080306061330075      1  0.47906     1   0.47906
 2 01080306061330078      1  0.56074     2   0.23230
 3 010803000301238868-1   1  0.19633     3   0.02132         2.75  0.83     2.84
 4 010803000301235585-2   0  0.94654     3   0.02944         2.30  1.14     1.84
 5 010803000301235768-1   0  0.86223     3   0.05371         1.72  1.01     1.10
 6 010803000301235607-2   0  0.73674     3   0.11746         1.23  0.90     1.72
 7 010803000301238349-1   0  0.84562     3   0.16743         1.01  0.83     2.14
 8 010803000301238838-1   0  0.89521     3   0.20658         0.88  0.79     1.26
 9 01080306910-1_2        0  0.76917     3   0.31169         0.67  0.76     0.43
10 010803000301236033-2   0  0.56700     3   0.61362         0.40  0.74     0.60
11 010803000301237450-3   0  0.61362     3   1.00000         0.19  0.72

Responses to items 1 to 3 are significant. p = 0.0213
Person parameter will be estimated based on the subscore of items 4 - 11

Rank correlation between person estimate and next item =  0.79
```

```
Estimates of person parameter

Optimal SEM with 11 dichotomous items ~  0.603

   WML =  0.192   SEM =  0.719   Bias =  0.000
    ML =  0.108   SEM =  0.846   Bias =  0.000


Subscore from 4 to 11 = 0  Max = 8

   WML = -1.618    SEM =  0.705

============================================

  +-----------------+
  |                 |
  | Elev nr. 386356 |
  |                 |
  +-----------------+

Profil = 1

Score = 6 out of 22

                                PCM
Item                    Score  Thresholds
0108010420130            1  1    1.10
010801000301238556-1     1  2    1.34
010801000301234839-1     1  3    1.66
0108010410315            0  3    0.28
010801000301239196-1     0  3    2.39
010801000301234841-1     0  3    0.88
0108010410094            0  3    0.86
0108010420122            0  3    1.14
0108010420160            0  3    1.67
010801000301238872-1     0  3    1.06
0108010415120            0  3    1.49
010801000301238996-1     0  3    0.28
010801000301238353-1     0  3   -0.06
0108010420142            0  3    0.71
0108010420150            0  3   -0.08
0108010415129            0  3   -0.86
0108010410230009         0  3    0.08
010801000301238279-1     1  4   -1.06
0108010420027            0  4   -0.74
0108010440028            1  5   -0.94
0108010415182            0  5   -0.95
0108010410230021         1  6   -0.59

Prob(1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1 | Score = 6) =  0.000005245

Monte Carlo test of person fit: p =   0.071

                        Conditional probabilites       Person estimates
```

Item	Item score	cumulated score		WML	SEM	Next
1 0108010420130	1 0.12922	1	0.12922			
2 010801000301238556-1	1 0.10348	2	0.01058			
3 010801000301234839-1	1 0.07744	3	0.00048	3.34	0.78	0.28
4 0108010410315	0 0.74130	3	0.00431	2.02	0.99	2.39
5 010801000301239196-1	0 0.96157	3	0.00608	1.74	0.95	0.88
6 010801000301234841-1	0 0.84303	3	0.01474	1.27	0.86	0.86
7 0108010410094	0 0.84076	3	0.02840	0.92	0.80	1.14
8 0108010420122	0 0.87584	3	0.04286	0.70	0.76	1.67
9 0108010420160	0 0.92342	3	0.05344	0.58	0.73	1.06
10 010801000301238872-1	0 0.86638	3	0.07404	0.41	0.71	1.49
11 0108010415120	0 0.90964	3	0.08982	0.31	0.70	0.28
12 010801000301238996-1	0 0.74130	3	0.14047	0.10	0.69	-0.06
13 010801000301238353-1	0 0.66628	3	0.21700	-0.13	0.68	0.71
14 0108010420142	0 0.81773	3	0.26030	-0.24	0.67	-0.08
15 0108010420150	0 0.66181	3	0.33570	-0.43	0.67	-0.86
16 0108010415129	0 0.45998	3	0.40596	-0.67	0.67	0.08
17 0108010410230009	0 0.69992	3	0.40268	-0.79	0.67	-1.06
18 010801000301238279-1	1 0.59420	4	0.43484	-0.65	0.61	-0.74
19 0108010420027	0 0.49228	4	0.43432	-0.83	0.61	-0.94
20 0108010440028	1 0.56189	5	0.54022	-0.69	0.56	-0.95
21 0108010415182	0 0.43426	5	0.46926	-0.86	0.56	-0.59
22 0108010410230021	1 0.46926	6	1.00000	-0.71	0.52	

```
Responses to items 1 to 3 are significant. p = 0.0005
Person parameter will be estimated based on the subscore of items 4 - 22

Rank correlation between person estimate and next item =  0.80

Estimates of person parameter

Optimal SEM with 22 dichotomous items ~  0.426

   WML = -0.708   SEM =  0.522   Bias =  0.000
    ML = -0.750   SEM =  0.564   Bias =  0.000


Subscore from 4 to 22 = 3  Max = 19

   WML = -1.568    SEM =  0.661

==============================================

  +-----------------+
  |                 |
  | Elev nr. 341070 |
  |                 |
  +-----------------+

Profil = 2

Score = 9 out of 17
```

```
Item                   Score  Thresholds
0108020111016           1  1    0.46
0108020111022           1  2    1.53
010802000301238021-1    1  3    1.34
010802000301239468-1    0  3    4.05
0108020110231           0  3    2.78
010802000301234818-1    0  3    3.19
0108020110166-1         0  3    2.12
0108020110268           0  3    2.28
0108020111006           1  4    1.39
010802000301239338-1    1  5    2.98
010802000301239337-1    0  5    3.39
010802000301234958-1    0  5    3.90
010802000301234959-1    0  5    1.99
0108020110139-1         1  6    1.29
010802000301234951-1    1  7    2.05
010802000301234848-1    1  8    2.09
0108020111019           1  9    1.37

Prob(1,1,1,0,0,0,0,0,1,1,0,0,0,1,1,1,1 | Score = 9) =  0.00175769

Monte Carlo test of person fit: p =   0.907

                       Conditional probabilites       Person estimates
Item                   Item score  cumulated score         WML   SEM     Next
 1 0108020111016        1  0.88171     1   0.88171
 2 0108020111022        1  0.70934     2   0.61919
 3 010802000301238021-1 1  0.74940     3   0.44915         3.17  0.80     4.05
 4 010802000301239468-1 0  0.85686     3   0.46633         3.04  1.21     2.78
 5 0108020110231        0  0.61011     3   0.47697         2.44  1.05     3.19
 6 010802000301234818-1 0  0.70859     3   0.44246         2.21  0.94     2.12
 7 0108020110166-1      0  0.43564     3   0.31543         1.85  0.86     2.28
 8 0108020110268        0  0.47717     3   0.17868         1.63  0.81     1.39
 9 0108020111006        1  0.73979     4   0.24432         1.84  0.74     2.98
10 010802000301239338-1 1  0.33949     5   0.39772         2.20  0.70     3.39
11 010802000301239337-1 0  0.75146     5   0.36250         2.11  0.67     3.90
12 010802000301234958-1 0  0.83653     5   0.33098         2.06  0.66     1.99
13 010802000301234959-1 0  0.40099     5   0.14729         1.86  0.63     1.29
14 0108020110139-1      1  0.75857     6   0.21622         1.98  0.60     2.05
15 010802000301234951-1 1  0.58183     7   0.41198         2.15  0.57     2.09
16 010802000301234848-1 1  0.57371     8   0.74353         2.29  0.55     1.37
17 0108020111019        1  0.74353     9   1.00000         2.36  0.54

Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item =  0.65

Estimates of person parameter

Optimal SEM with 17 dichotomous items ~  0.485

   WML =  2.361   SEM =  0.536   Bias =  0.000
    ML =  2.370   SEM =  0.562   Bias =  0.000
```

```
============================================

  +-----------------+
  |                 |
  | Elev nr. 317854 |
  |                 |
  +-----------------+

Profil = 2

Score = 15 out of 32

Thresholds are PCM parameters

                                PCM
Item                   Score  Thresholds
0108020111018           1  1    1.05
010802000301234810-1    1  2    1.47
0108020111006           1  3    1.39
010802000301239468-1    0  3    4.05
0108020110231           0  3    2.78
010802000301234818-1    0  3    3.19
0108020110166-1         0  3    2.12
0108020110268           0  3    2.28
010802000301238021-1    1  4    1.34
010802000301239338-1    1  5    2.98
010802000301239337-1    0  5    3.39
010802000301234958-1    0  5    3.90
010802000301234959-1    0  5    1.99
0108020110139-1         1  6    1.29
010802000301234951-1    1  7    2.05
010802000301234848-1    1  8    2.09
0108020111019           1  9    1.37
0108020111022           0  9    1.53
0108020111026           1 10    0.87
010802000301234814-1    0 10    2.15
010802000301234953-1    1 11    1.42
0108020110266           1 12    0.58
010802000301234788-1    1 13    1.35
010802000301234941-1    0 13    1.39
0108020110150-1         1 14    1.14
0108020110083-1         0 14    1.30
0108020111016           1 15    0.46
010802000301234966-1    0 15    1.37
0108020115072           0 15    1.38
0108020111032           0 15    0.29
0108020110027-1         0 15    1.50
0108020111024           0 15    0.42

Prob(1,1,1,0,0,0,0,0,1,1,0,0,0,1,1,1,1,0,1,0,1,1,1,0,1,0,1,0,0,0,0,0 | Score = 15) =  0.00000001464

Monte Carlo test of person fit: p =   0.397

                        Conditional probabilites       Person estimates
```

Item	Item score	cumulated score		WML	SEM	Next
1 0108020111018	1 0.62703	1	0.62703			
2 010802000301234810-1	1 0.52247	2	0.31861			
3 0108020111006	1 0.54391	3	0.16247	3.27	0.78	4.05
4 010802000301239468-1	0 0.92793	3	0.18287	3.10	1.19	2.78
5 0108020110231	0 0.77922	3	0.24135	2.51	1.03	3.19
6 010802000301234818-1	0 0.84327	3	0.27372	2.28	0.92	2.12
7 0108020110166-1	0 0.64253	3	0.32945	1.93	0.85	2.28
8 0108020110268	0 0.67847	3	0.34664	1.70	0.79	1.34
9 010802000301238021-1	1 0.55621	4	0.29032	1.89	0.73	2.98
10 010802000301239338-1	1 0.18707	5	0.16051	2.24	0.69	3.39
11 010802000301239337-1	0 0.86902	5	0.18088	2.15	0.67	3.90
12 010802000301234958-1	0 0.91700	5	0.19327	2.09	0.65	1.99
13 010802000301234959-1	0 0.61051	5	0.24811	1.90	0.63	1.29
14 0108020110139-1	1 0.56818	6	0.18805	2.01	0.60	2.05
15 010802000301234951-1	1 0.37336	7	0.10679	2.17	0.57	2.09
16 010802000301234848-1	1 0.36591	8	0.04952	2.31	0.55	1.37
17 0108020111019	1 0.54866	9	0.02520	2.38	0.53	1.53
18 0108020111022	0 0.49337	9	0.05363	2.20	0.51	0.87
19 0108020111026	1 0.67088	10	0.03241	2.25	0.50	2.15
20 010802000301234814-1	0 0.65012	10	0.05311	2.13	0.49	1.42
21 010802000301234953-1	1 0.53513	11	0.02447	2.20	0.48	0.58
22 0108020110266	1 0.73230	12	0.01363	2.23	0.47	1.35
23 010802000301234788-1	1 0.55381	13	0.00428	2.29	0.46	1.39
24 010802000301234941-1	0 0.45808	13	0.01168	2.15	0.45	1.14
25 0108020110150-1	1 0.60603	14	0.00317	2.20	0.44	1.30
26 0108020110083-1	0 0.43494	14	0.00932	2.07	0.43	0.46
27 0108020111016	1 0.75625	15	0.00252	2.10	0.42	1.37
28 010802000301234966-1	0 0.45218	15	0.00729	1.98	0.41	1.38
29 0108020115072	0 0.45581	15	0.01951	1.88	0.40	0.29
30 0108020111032	0 0.21346	15	0.10768	1.75	0.40	1.50
31 0108020110027-1	0 0.48517	15	0.23613	1.67	0.39	0.42
32 0108020111024	0 0.23613	15	1.00000	1.56	0.38	

Nothing unexpected during the start or end of the test

Rank correlation between person estimate and next item = 0.64

Estimates of person parameter

Optimal SEM with 32 dichotomous items ~ 0.354

WML = 1.556 SEM = 0.384 Bias = 0.000
ML = 1.555 SEM = 0.394 Bias = 0.000

==========

C

Item-parametre

Dette bilag viser itemsværhedsgraderne ifølge 2017-analysen. Beta er itemmets sværhedsgrad, t.1-t.7 er thresholds/step-parametrene for polytome items. se.t.1-se.t.7 er standardmålefejlen for parametrene. Derefter vises DNT's estimater af itemsværhedsgraderne. For partial credit items er værdierne angivet som såkaldte ucentraliserede thresholdværdier der ikke har en let fortolkning (det er skæringspunkterne mellem de såkaldte itemkarakteristiske kurver). Værdierne kan omregnes til såkaldte Thurstonian thresholds hvorved de er lettere at fortolke. Dette har ikke været nødvendigt for analyserne i denne rapport.

Tabel C.1: Item-parametre

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
01080100030123478 9-1	1,447	1,447							1,202	1,202							0,041						
01080100030123480 2-1	1,480	1,480							1,563	1,563							0,032						
01080100030123480 4-1	1,165	1,165							1,505	1,505							0,035						
01080100030123480 6-1	0,204	0,204							0,628	0,628							0,037						
01080100030123480 7-1	0,901	0,901							1,566	1,566							0,031						
01080100030123481 5-1	0,898	0,898							0,800	0,800							0,045						
01080100030123481 7-1	1,506	1,506							1,346	1,346							0,042						
01080100030123482 0-1	0,535	0,535							1,552	1,552							0,034						
01080100030123482 3-1	1,235	1,235							1,079	1,079							0,038						
01080100030123482 4-1	0,581	0,581							1,685	1,685							0,027						
01080100030123482 5-1	1,164	1,164							1,288	1,288							0,043						
01080100030123482 6-1	1,068	1,068							1,396	1,396							0,039						
01080100030123482 8-1	0,776	0,776							1,912	1,912							0,026						
01080100030123482 9-1	1,138	1,138							1,634	1,634							0,029						
01080100030123483 0-1	-1,701	-1,701							-1,449	-1,449							0,033						
01080100030123483 1-1	1,010	1,010							0,811	0,811							0,044						
01080100030123483 2-1	0,763	0,763							0,554	0,554							0,028						
01080100030123483 3-1	0,711	0,711							1,937	1,937							0,026						
01080100030123483 5-1	-0,655	-0,655							-0,286	-0,286							0,028						
01080100030123483 6-1	1,179	1,179							0,875	0,875							0,046						
01080100030123483 7-1	1,240	1,240							2,263	2,263							0,026						
01080100030123483 8-1	1,046	1,046							0,930	0,930							0,036						
01080100030123483 9-1	1,655	1,655							1,495	1,495							0,035						
01080100030123484 0-1	1,261	1,261							1,155	1,155							0,043						
01080100030123484 1-1	0,879	0,879							2,248	2,248							0,026						
01080100030123484 2-1	0,973	0,973							0,879	0,879							0,045						
01080100030123484 3-1	1,115	1,115							1,463	1,463							0,036						
01080100030123606 4-1	-1,931	-1,931							-1,876	-1,876							0,042						
01080100030123606 8-1	1,939	1,939							1,649	1,649							0,028						
01080100030123607 3-1	1,871	1,871							1,532	1,532							0,034						
01080100030123827 7-1	-0,470	-0,470							-0,680	-0,680							0,031						
01080100030123827 8-1	-1,662	-1,662							-1,064	-1,064							0,034						
01080100030123827 9-1	-1,063	-1,063							-0,779	-0,779							0,035						
01080100030123828 1-1	0,764	0,764							0,745	0,745							0,041						
01080100030123828 2-1	0,708	0,708							0,326	0,326							0,034						
01080100030123828 3-1	-0,590	-0,590							-0,944	-0,944							0,033						
01080100030123828 5-1	-0,950	-0,950							-0,576	-0,576							0,029						
01080100030123835 3-1	-0,062	-0,062							0,232	0,232							0,033						
01080100030123855 5-1	0,300	0,300							0,479	0,479							0,028						
01080100030123855 6-1	1,344	1,344							1,075	1,075							0,038						
01080100030123860 7-1	0,930	0,930							0,591	0,591							0,029						
01080100030123883 5-1	1,278	1,278							0,775	0,775							0,044						
01080100030123883 6-1	1,335	1,335							0,862	0,862							0,047						
01080100030123883 7-1	0,254	0,254							-0,178	-0,178							0,029						
01080100030123883 9-1	1,065	1,065							0,735	0,735							0,042						
01080100030123886 2-1	1,802	1,802							1,248	1,248							0,044						
01080100030123887 2-1	1,062	1,062							1,003	1,003							0,034						
01080100030123887 3-1	1,258	1,258							0,843	0,843							0,046						
01080100030123892 0-1	1,710	1,710							1,299	1,299							0,045						
01080100030123892 3-1	1,497	1,497							1,306	1,306							0,044						
01080100030123892 5-1	1,346	1,346							0,743	0,743							0,043						
01080100030123892 7-1	0,466	0,466							0,463	0,463							0,029						
01080100030123899 4-1	-1,044	-1,044							-0,652	-0,652							0,032						
01080100030123899 5-1	-0,265	-0,265							-0,216	-0,216							0,027						
01080100030123899 6-1	0,279	0,279							0,448	0,448							0,028						
01080100030123899 7-1	1,837	1,837							1,445	1,445							0,036						
01080100030123899 9-1	0,940	0,940							0,836	0,836							0,045						
01080100030123900 0-1	1,452	1,452							1,367	1,367							0,041						
01080100030123919 5-1	1,319	1,319							0,760	0,760							0,044						
01080100030123919 6-1	2,388	2,388							2,371	2,371							0,027						
01080100030123919 7-1	-0,117	-0,117							-0,073	-0,073							0,029						
01080100030123919 9-1	-1,510	-1,510							-1,527	-1,527							0,036						
01080100030123920 0-1	-0,956	-0,956							-0,834	-0,834							0,031						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108010003012392 35-1	-0,109	-0,109							-0,295	-0,295							0,028						
0108010410080	-2,657	-2,657							-2,399	-2,399							0,043						
0108010410083	-0,039	-0,039							0,433	0,433							0,028						
0108010410084	1,028	1,028							1,081	1,081							0,037						
0108010410088	1,297	1,297							0,801	0,801							0,046						
0108010410093	-1,356	-1,356							-0,793	-0,793							0,035						
0108010410094	0,863	0,863							1,832	1,832							0,026						
0108010410096	-1,819	-1,819							-0,680	-0,680							0,034						
0108010410097	-1,850	-1,850							-1,415	-1,415							0,034						
0108010410098	-1,742	-1,742							-0,938	-0,938							0,034						
0108010410110-1	-0,888	-0,888							-0,612	-0,612							0,032						
0108010410145-1	1,218	1,218							0,793	0,793							0,045						
0108010410155-1	0,345	0,345							-0,001	-0,001							0,027						
0108010410186-1	-3,509	-3,509							-4,105	-4,105							0,051						
0108010410187-1	-1,086	-1,086							-0,622	-0,622							0,033						
0108010410230005	0,128	0,128							0,305	0,305							0,033						
0108010410230008	-0,287	-0,287							0,193	0,193							0,033						
0108010410230009	0,084	0,084							-0,567	-0,567							0,029						
0108010410230012	1,023	1,023							0,733	0,733							0,041						
0108010410230013	0,422	0,422							0,116	0,116							0,032						
0108010410230014	0,070	0,070							-0,412	-0,412							0,028						
0108010410230017	1,176	1,176							0,224	0,224							0,036						
0108010410230018	-0,716	-0,716							-0,067	-0,067							0,029						
0108010410230019	0,742	0,742							0,271	0,271							0,033						
0108010410230020	-0,178	-0,178							-0,526	-0,526							0,028						
0108010410230021	-0,592	-0,592							-0,733	-0,733							0,033						
0108010410230022	0,115	0,115							0,124	0,124							0,031						
0108010410230023	0,953	0,953							0,257	0,257							0,034						
0108010410230024	1,268	1,268							0,851	0,851							0,047						
0108010410230025	-1,556	-1,556							-1,257	-1,257							0,037						
0108010410230028	0,207	0,207							-0,180	-0,180							0,028						
0108010410230029	-0,977	-0,977							-0,570	-0,570							0,029						
0108010410230030	-0,278	-0,278							-0,163	-0,163							0,028						
0108010410230031	1,157	1,157							0,550	0,550							0,030						
0108010410230032	-0,899	-0,899							-0,713	-0,713							0,032						
0108010410230034	1,697	1,697							0,668	0,668							0,042						
0108010410230035	1,156	1,156							1,135	1,135							0,043						
0108010410230037	1,492	1,492							0,901	0,901							0,039						
0108010410230038	0,518	0,518							0,174	0,174							0,032						
0108010410230039	-1,191	-1,191							-0,437	-0,437							0,029						
0108010410230041	-0,080	-0,080							-0,376	-0,376							0,030						
0108010410230042	-1,694	-1,694							-1,792	-1,792							0,041						
0108010410230045	-0,907	-0,907							-1,474	-1,474							0,034						
0108010410230046	-0,223	-0,223							-0,021	-0,021							0,027						
0108010410230047	1,490	1,490							1,047	1,047							0,037						
0108010410311	-2,340	-2,340							-4,135	-4,135							0,049						
0108010410315	0,279	0,279							3,044	3,044							0,028						
0108010410316	-1,212	-1,212							-2,333	-2,333							0,042						
0108010410320	0,039	0,039							-0,256	-0,256							0,027						
0108010410325	-1,192	-1,192							-0,871	-0,871							0,031						
0108010410327	-1,163	-1,163							-1,672	-1,672							0,041						
0108010410328	1,449	1,449							0,844	0,844							0,047						
0108010410333	-0,874	-0,874							-0,767	-0,767							0,034						
0108010410335	-1,671	-1,671							-1,790	-1,790							0,041						
0108010410337	-1,943	-1,943							-1,932	-1,932							0,041						
0108010410339	-1,788	-1,788							-1,584	-1,584							0,042						
0108010410340	-2,794	-2,794							-2,741	-2,741							0,042						
0108010410343	-1,889	-1,889							-2,434	-2,434							0,043						
0108010410344	-2,666	-2,666							-3,694	-3,694							0,049						
0108010410350	1,099	1,099							1,464	1,464							0,035						
0108010410351	-0,692	-0,692							-0,283	-0,283							0,028						
0108010410357	-0,661	-0,661							0,193	0,193							0,034						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108010410358	-1,861	-1,861							-1,005	-1,005							0,035						
0108010410366	0,072	0,072							0,569	0,569							0,028						
0108010410368	0,937	0,937							1,318	1,318							0,043						
0108010410369	-2,473	-2,473							-2,610	-2,610							0,037						
0108010410372	0,855	0,855							0,917	0,917							0,036						
0108010410373	0,844	0,844							0,616	0,616							0,037						
0108010410376	-1,856	-1,856							-1,752	-1,752							0,040						
0108010410377	1,231	1,231							0,805	0,805							0,045						
0108010410378	1,141	1,141							1,009	1,009							0,035						
0108010410379	2,174	2,174							1,807	1,807							0,026						
0108010410384	0,890	0,890							0,931	0,931							0,036						
0108010410385	-1,935	-1,935							-1,614	-1,614							0,044						
0108010410388	1,171	1,171							1,049	1,049							0,036						
0108010410392	-0,117	-0,117							0,512	0,512							0,028						
0108010410393	0,114	0,114							0,403	0,403							0,028						
0108010410395	-1,198	-1,198							-1,192	-1,192							0,034						
0108010410397	-0,120	-0,120							0,356	0,356							0,034						
0108010410398	0,344	0,344							0,505	0,505							0,028						
0108010410399	1,465	1,465							1,160	1,160							0,043						
0108010410400-1	-0,637	-0,637							-0,279	-0,279							0,027						
0108010410401-1	-0,770	-0,770							-1,020	-1,020							0,033						
0108010410402-1	-1,609	-1,609							-0,989	-0,989							0,034						
0108010410405-1	1,175	1,175							1,143	1,143							0,043						
0108010410406-1	-0,922	-0,922							-0,912	-0,912							0,031						
0108010410407-1	0,860	0,860							1,524	1,524							0,034						
0108010410408-1	-2,732	-2,732							-1,247	-1,247							0,042						
0108010410410-1	-1,825	-1,825							-2,504	-2,504							0,040						
0108010410411-1	-1,087	-1,087							-0,467	-0,467							0,028						
0108010410412-1	1,792	1,792							1,282	1,282							0,046						
0108010410413-1	-0,070	-0,070							-0,108	-0,108							0,031						
0108010410414-1	-0,720	-0,720							-0,178	-0,178							0,028						
0108010415102	-2,238	-2,238							-2,133	-2,133							0,036						
0108010415103	-2,341	-2,341							-2,771	-2,771							0,041						
0108010415109	1,427	1,427							1,266	1,266							0,043						
0108010415111	0,449	0,449							0,192	0,192							0,033						
0108010415113	1,329	1,329							1,294	1,294							0,044						
0108010415117	-1,187	-1,187							-0,648	-0,648							0,033						
0108010415118	1,104	1,104							1,095	1,095							0,038						
0108010415119	0,364	0,364							0,237	0,237							0,033						
0108010415120	1,490	1,490							0,702	0,702							0,042						
0108010415122	0,191	0,191							0,322	0,322							0,033						
0108010415124	0,972	0,972							1,713	1,713							0,026						
0108010415129	-0,856	-0,856							-0,372	-0,372							0,030						
0108010415130	0,111	0,111							0,132	0,132							0,031						
0108010415132	-0,901	-0,901							-0,395	-0,395							0,030						
0108010415133	1,097	1,097							0,566	0,566							0,029						
0108010415134	0,049	0,049							0,380	0,380							0,035						
0108010415135	-0,235	-0,235							0,347	0,347							0,034						
0108010415139	-2,257	-2,257							-0,906	-0,906							0,035						
0108010415140	-0,225	-0,225							-0,437	-0,437							0,028						
0108010415145	0,012	0,012							0,409	0,409							0,028						
0108010415151	1,972	1,972							1,370	1,370							0,043						
0108010415153	1,499	1,499							1,381	1,381							0,040						
0108010415157	-1,183	-1,183							-1,140	-1,140							0,033						
0108010415158	0,590	0,590							0,951	0,951							0,034						
0108010415159	-0,003	-0,003							0,050	0,050							0,027						
0108010415160	0,611	0,611							0,090	0,090							0,028						
0108010415164	-2,144	-2,144							-1,848	-1,848							0,041						
0108010415165	-0,336	-0,336							0,893	0,893							0,046						
0108010415167	1,452	1,452							1,160	1,160							0,043						
0108010415169	-2,471	-2,471							-1,204	-1,204							0,039						
0108010415171	-2,703	-2,703							-2,007	-2,007							0,038						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108010415173	0,836	0,836							0,714	0,714							0,040						
0108010415175	-0,646	-0,646							-0,778	-0,778							0,035						
0108010415178	-1,160	-1,160							-0,786	-0,786							0,035						
0108010415179	-2,373	-2,373							-1,958	-1,958							0,040						
0108010415180	2,493	2,493							1,981	1,981							0,027						
0108010415182	-0,953	-0,953							-0,629	-0,629							0,032						
0108010415183	0,798	0,798							1,026	1,026							0,035						
0108010415186	0,162	0,162							0,674	0,674							0,038						
0108010415190	-2,469	-2,469							-1,984	-1,984							0,038						
0108010415193	-1,814	-1,814							-1,641	-1,641							0,043						
0108010415194	-1,717	-1,717							-1,322	-1,322							0,036						
0108010420002	-0,742	-0,742							-1,000	-1,000							0,034						
0108010420003	-0,410	-0,410							-0,103	-0,103							0,031						
0108010420010	-1,038	-1,038							-0,352	-0,352							0,030						
0108010420012	-1,614	-1,614							-1,121	-1,121							0,033						
0108010420013	0,085	0,085							0,516	0,516							0,028						
0108010420014	0,392	0,392							0,664	0,664							0,037						
0108010420015	0,807	0,807							2,151	2,151							0,026						
0108010420016	-2,703	-2,703							-3,657	-3,657							0,045						
0108010420017	0,209	0,209							1,235	1,235							0,041						
0108010420018	0,473	0,473							0,583	0,583							0,028						
0108010420019	0,251	0,251							0,398	0,398							0,036						
0108010420021	-1,158	-1,158							-1,922	-1,922							0,042						
0108010420023	-0,297	-0,297							-0,472	-0,472							0,028						
0108010420024	0,049	0,049							0,074	0,074							0,027						
0108010420027	-0,736	-0,736							-0,617	-0,617							0,032						
0108010420028	1,334	1,334							0,897	0,897							0,046						
0108010420029	-1,699	-1,699							-1,340	-1,340							0,035						
0108010420030	-0,674	-0,674							-0,325	-0,325							0,028						
0108010420031	0,282	0,282							0,821	0,821							0,044						
0108010420032	-2,177	-2,177							-2,525	-2,525							0,039						
0108010420033	-0,365	-0,365							-0,466	-0,466							0,027						
0108010420034	-2,100	-2,100							-1,625	-1,625							0,045						
0108010420040	0,468	0,468							0,394	0,394							0,036						
0108010420041	-1,728	-1,728							-1,242	-1,242							0,037						
0108010420042	-0,926	-0,926							-0,769	-0,769							0,035						
0108010420043	0,013	0,013							0,464	0,464							0,028						
0108010420044	-1,984	-1,984							-1,272	-1,272							0,038						
0108010420045	-0,734	-0,734							-0,052	-0,052							0,029						
0108010420046	-3,038	-3,038							-4,128	-4,128							0,050						
0108010420048	-0,842	-0,842							0,062	0,062							0,028						
0108010420049	-0,093	-0,093							0,356	0,356							0,034						
0108010420050	-0,683	-0,683							-0,515	-0,515							0,027						
0108010420053	0,033	0,033							0,171	0,171							0,032						
0108010420054	0,837	0,837							0,853	0,853							0,046						
0108010420056	1,115	1,115							1,259	1,259							0,043						
0108010420058	-0,553	-0,553							-0,015	-0,015							0,028						
0108010420059	1,210	1,210							0,462	0,462							0,031						
0108010420060	-1,888	-1,888							-1,573	-1,573							0,042						
0108010420061	0,985	0,985							0,294	0,294							0,034						
0108010420064	-1,797	-1,797							-1,351	-1,351							0,035						
0108010420066	0,261	0,261							0,393	0,393							0,036						
0108010420067	-1,936	-1,936							-1,086	-1,086							0,035						
0108010420068	0,533	0,533							0,904	0,904							0,037						
0108010420070	-0,040	-0,040							0,498	0,498							0,028						
0108010420071	-1,428	-1,428							-1,619	-1,619							0,045						
0108010420073	-1,194	-1,194							-0,049	-0,049							0,031						
0108010420074	-0,675	-0,675							0,470	0,470							0,029						
0108010420079	1,014	1,014							0,920	0,920							0,036						
0108010420084	-1,589	-1,589							-0,544	-0,544							0,030						
0108010420087	1,322	1,322							1,026	1,026							0,036						
0108010420088	1,134	1,134							0,802	0,802							0,045						
0108010420092	0,548	0,548							0,020	0,020							0,028						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108010420094	0,421	0,421							0,239	0,239							0,033						
0108010420095	1,880	1,880							1,716	1,716							0,026						
0108010420100	0,396	0,396							0,667	0,667							0,037						
0108010420101	-2,020	-2,020							-1,920	-1,920							0,041						
0108010420102	0,619	0,619							0,245	0,245							0,033						
0108010420103	1,520	1,520							1,152	1,152							0,044						
0108010420104	-0,759	-0,759							-0,384	-0,384							0,030						
0108010420105	1,529	1,529							0,849	0,849							0,048						
0108010420106	1,091	1,091							0,635	0,635							0,038						
0108010420109	1,251	1,251							1,102	1,102							0,044						
0108010420112	0,987	0,987							0,479	0,479							0,030						
0108010420113	0,835	0,835							0,534	0,534							0,029						
0108010420114	1,138	1,138							0,696	0,696							0,040						
0108010420116	-0,891	-0,891							-0,739	-0,739							0,033						
0108010420120	1,700	1,700							0,700	0,700							0,043						
0108010420122	1,144	1,144							1,465	1,465							0,035						
0108010420126	-2,060	-2,060							-2,616	-2,616							0,036						
0108010420129	1,514	1,514							1,085	1,085							0,039						
0108010420130	1,099	1,099							0,463	0,463							0,030						
0108010420131	-0,004	-0,004							-0,057	-0,057							0,029						
0108010420132	1,472	1,472							0,877	0,877							0,047						
0108010420136	0,245	0,245							0,008	0,008							0,027						
0108010420138	1,454	1,454							0,870	0,870							0,047						
0108010420140	1,711	1,711							0,760	0,760							0,045						
0108010420141	0,061	0,061							-0,072	-0,072							0,028						
0108010420142	0,706	0,706							0,068	0,068							0,028						
0108010420147	0,906	0,906							0,868	0,868							0,046						
0108010420149	0,814	0,814							0,458	0,458							0,029						
0108010420150	-0,081	-0,081							-0,134	-0,134							0,030						
0108010420151	0,985	0,985							0,402	0,402							0,029						
0108010420152	0,877	0,877							0,817	0,817							0,045						
0108010420153	1,368	1,368							1,097	1,097							0,039						
0108010420154	-0,093	-0,093							0,550	0,550							0,028						
0108010420155	0,810	0,810							0,580	0,580							0,029						
0108010420156	1,232	1,232							0,614	0,614							0,038						
0108010420157	1,232	1,232							0,499	0,499							0,030						
0108010420160	1,667	1,667							1,249	1,249							0,044						
0108010420161	-0,888	-0,888							-0,227	-0,227							0,028						
0108010420162	-0,140	-0,140							0,547	0,547							0,028						
0108010440001	0,615	0,615							0,167	0,167							0,032						
0108010440006	1,811	1,811							1,020	1,020							0,038						
0108010440007	1,118	1,118							0,930	0,930							0,036						
0108010440010	0,024	0,024							-0,044	-0,044							0,029						
0108010440013	-0,920	-0,920							-0,057	-0,057							0,030						
0108010440014	-1,065	-1,065							-0,923	-0,923							0,032						
0108010440019	-2,370	-2,370							-1,622	-1,622							0,045						
0108010440021	-0,600	-0,600							-1,043	-1,043							0,033						
0108010440022	-1,984	-1,984							-2,187	-2,187							0,037						
0108010440025	0,289	0,289							0,109	0,109							0,031						
0108010440027	-1,934	-1,934							-1,945	-1,945							0,040						
0108010440028	-0,939	-0,939							-0,807	-0,807							0,033						
0108010440031	-0,474	-0,474							-0,102	-0,102							0,032						
0108010440034	1,774	1,774							1,104	1,104							0,047						
0108010440036	-0,229	-0,229							-0,143	-0,143							0,029						
0108010440037	0,036	0,036							0,080	0,080							0,027						
0108010440040	0,105	0,105							0,102	0,102							0,031						
0108010440042	-0,680	-0,680							-0,574	-0,574							0,028						
0108010440045	1,360	1,360							0,805	0,805							-10,823						
01080200030123478 8-1	1,347	1,347							2,183	2,183							0,019						
01080200030123481 0-1	1,469	1,469							2,569	2,569							0,012						
01080200030123481 1-1	0,501	0,501							0,805	0,805							0,032						
01080200030123481 3-1	-0,628	-0,628							-0,639	-0,639							0,046						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
010802000301234814-1	2,153	2,153							2,278	2,278							0,017						
010802000301234816-1	0,829	0,829							1,667	1,667							0,024						
010802000301234818-1	3,188	3,188							4,058	4,058							0,017						
010802000301234848-1	2,085	2,085							2,646	2,646							0,014						
010802000301234855-1	0,290	0,290							0,990	0,990							0,034						
010802000301234876-1	2,963	2,963							4,717	4,717							0,022						
010802000301234881-1	0,269	0,269							0,539	0,539							0,032						
010802000301234885-1	4,002	4,002							7,961	7,961							0,044						
010802000301234941-1	1,394	1,394							2,150	2,150							0,019						
010802000301234943-1	0,776	0,776							1,664	1,664							0,025						
010802000301234946-1	0,283	0,283							1,353	1,353							0,021						
010802000301234951-1	2,054	2,054							2,651	2,651							0,013						
010802000301234952-1	1,087	1,087							1,497	1,497							0,016						
010802000301234953-1	1,420	1,420							2,228	2,228							0,017						
010802000301234955-1	-0,991	-0,991							-1,138	-1,138							0,046						
010802000301234956-1	0,412	0,412							1,305	1,305							0,022						
010802000301234958-1	3,897	3,897							4,046	4,046							0,020						
010802000301234959-1	1,988	1,988							2,793	2,793							0,012						
010802000301234962-1	-0,786	-0,786							-0,422	-0,422							0,046						
010802000301234965-1	1,072	1,072							1,610	1,610							0,024						
010802000301234966-1	1,371	1,371							1,973	1,973							0,013						
010802000301234973-1	3,930	3,930							6,117	6,117							0,033						
010802000301237730-1	-2,375	-2,375							-3,038	-3,038							0,054						
010802000301237980-1	-2,550	-2,550							-5,039	-5,039							0,055						
010802000301237981-1	-1,877	-1,877							-4,589	-4,589							0,055						
010802000301237982-1	-2,442	-2,442							-3,455	-3,455							0,054						
010802000301237983-1	-2,626	-2,626							-3,190	-3,190							0,054						
010802000301237984-1	-1,694	-1,694							-4,942	-4,942							0,055						
010802000301237986-1	-2,245	-2,245							-4,503	-4,503							0,054						
010802000301238020-1	1,574	1,574							1,630	1,630							0,026						
010802000301238021-1	1,338	1,338							3,045	3,045							0,013						
010802000301238024-1	-1,274	-1,274							-1,150	-1,150							0,047						
010802000301239337-1	3,394	3,394							3,706	3,706							0,016						
010802000301239338-1	2,979	2,979							3,720	3,720							0,015						
010802000301239438-1	0,512	0,512							0,704	0,704							0,029						
010802000301239468-1	4,049	4,049							5,044	5,044							0,024						
010802000301239469-1	3,640	3,640							4,656	4,656							0,023						
0108020110002-1	-1,557	-1,557							0,113	0,113							0,046						
0108020110003-1	-0,628	-0,628							0,459	0,459							0,038						
0108020110004-1	-0,876	-0,876							-1,286	-1,286							0,050						
0108020110005-1	-0,631	-0,631							-0,353	-0,353							0,044						
0108020110006-1	-0,392	-0,392							0,904	0,904							0,037						
0108020110007-1	-0,893	-0,893							-1,183	-1,183							0,047						
0108020110012-1	-0,642	-0,642							-1,253	-1,253							0,049						
0108020110013-1	-0,441	-0,441							-1,016	-1,016							0,048						
0108020110018-1	0,022	0,022							0,847	0,847							0,035						
0108020110023-1	-0,346	-0,346							0,105	0,105							0,042						
0108020110024-1	0,532	0,532							0,732	0,732							0,030						
0108020110025-1	-0,907	-0,907							-1,732	-1,732							0,054						
0108020110027-1	1,498	1,498							1,827	1,827							0,020						
0108020110028-1	-0,381	-0,381							-2,275	-2,275							0,053						
0108020110029-1	0,016	0,016							0,361	0,361							0,042						
0108020110030-1	-0,156	-0,156							-2,457	-2,457							0,054						
0108020110036-1	0,291	0,291							0,904	0,904							0,035						
0108020110037-1	-0,593	-0,593							-2,003	-2,003							0,053						
0108020110038-1	-0,046	-0,046							0,269	0,269							0,039						
0108020110039-1	0,058	0,058							0,462	0,462							0,036						
0108020110046-1	-0,682	-0,682							-2,423	-2,423							0,053						
0108020110049-1	-0,789	-0,789							-0,932	-0,932							0,051						
0108020110050-1	-0,317	-0,317							-0,416	-0,416							0,045						
0108020110051-1	-0,571	-0,571							-1,725	-1,725							0,054						
0108020110052-1	-0,840	-0,840							-1,545	-1,545							0,051						
0108020110053-1	-0,838	-0,838							-0,606	-0,606							0,046						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108020110054-1	-0,814	-0,814							-0,267	-0,267							0,046						
0108020110059-1	0,433	0,433							0,307	0,307							0,040						
0108020110062-1	-0,142	-0,142							-1,803	-1,803							0,053						
0108020110064-1	-0,690	-0,690							-0,934	-0,934							0,051						
0108020110065-1	-0,340	-0,340							-0,250	-0,250							0,045						
0108020110066-1	-0,065	-0,065							-0,292	-0,292							0,045						
0108020110068-1	-0,733	-0,733							-1,330	-1,330							0,050						
0108020110069-1	1,307	1,307							1,207	1,207							0,026						
0108020110074-1	0,452	0,452							0,979	0,979							0,034						
0108020110075-1	-0,545	-0,545							-0,512	-0,512							0,045						
0108020110076-1	0,161	0,161							0,457	0,457							0,037						
0108020110079-1	0,217	0,217							0,916	0,916							0,036						
0108020110081-1	0,398	0,398							1,549	1,549							0,017						
0108020110082-1	0,130	0,130							0,384	0,384							0,041						
0108020110083-1	1,303	1,303							2,139	2,139							0,019						
0108020110084-1	-0,451	-0,451							-0,101	-0,101							0,043						
0108020110085-1	0,048	0,048							1,009	1,009							0,034						
0108020110086-1	-0,646	-0,646							-1,120	-1,120							0,045						
0108020110088-1	0,429	0,429							0,892	0,892							0,038						
0108020110090-1	-2,034	-2,034							-1,334	-1,334							0,052						
0108020110096-1	-1,256	-1,256							-1,647	-1,647							0,054						
0108020110097-1	-1,038	-1,038							-0,468	-0,468							0,046						
0108020110100-1	-0,074	-0,074							0,078	0,078							0,041						
0108020110102-1	-0,171	-0,171							0,128	0,128							0,040						
0108020110103-1	-0,277	-0,277							-0,825	-0,825							0,052						
0108020110104-1	-0,184	-0,184							0,171	0,171							0,037						
0108020110105-1	-0,956	-0,956							0,592	0,592							0,034						
0108020110106-1	-0,624	-0,624							-0,795	-0,795							0,052						
0108020110108-1	-0,254	-0,254							-0,848	-0,848							0,051						
0108020110111-1	-1,311	-1,311							-1,828	-1,828							0,053						
0108020110113-1	0,052	0,052							-1,618	-1,618							0,054						
0108020110114-1	-0,138	-0,138							-1,762	-1,762							0,054						
0108020110130-1	-0,631	-0,631							-0,707	-0,707							0,048						
0108020110131-1	0,213	0,213							0,877	0,877							0,037						
0108020110134-1	-0,098	-0,098							-1,015	-1,015							0,048						
0108020110136-1	-0,484	-0,484							-2,952	-2,952							0,054						
0108020110137-1	0,776	0,776							1,035	1,035							0,033						
0108020110139-1	1,291	1,291							2,696	2,696							0,013						
0108020110141-1	-0,434	-0,434							-2,324	-2,324							0,053						
0108020110144-1	0,391	0,391							0,916	0,916							0,035						
0108020110147-1	-0,178	-0,178							-1,419	-1,419							0,049						
0108020110148-1	-0,362	-0,362							-2,121	-2,121							0,053						
0108020110149-1	0,079	0,079							-0,808	-0,808							0,052						
0108020110150-1	1,141	1,141							2,149	2,149							0,019						
0108020110155-1	-0,190	-0,190							0,266	0,266							0,038						
0108020110156-1	-0,118	-0,118							0,864	0,864							0,037						
0108020110162-1	1,152	1,152							0,789	0,789							0,033						
0108020110164-1	-0,802	-0,802							1,045	1,045							0,036						
0108020110166-1	2,121	2,121							3,463	3,463							0,013						
0108020110171-1	-1,333	-1,333							-2,574	-2,574							0,052						
0108020110172-1	-1,375	-1,375							-2,635	-2,635							0,051						
0108020110176-1	-0,857	-0,857							-0,755	-0,755							0,051						
0108020110177-1	-1,454	-1,454							-3,964	-3,964							0,053						
0108020110178-1	-1,846	-1,846							-2,083	-2,083							0,054						
0108020110180-1	-1,848	-1,848							-1,663	-1,663							0,054						
0108020110185-1	-0,574	-0,574							-2,716	-2,716							0,051						
0108020110186-1	-1,249	-1,249							-2,557	-2,557							0,052						
0108020110187-1	-0,643	-0,643							-2,853	-2,853							0,054						
0108020110188-1	-1,197	-1,197							-2,447	-2,447							0,053						
0108020110189-1	-1,585	-1,585							-0,025	-0,025							0,046						
0108020110190-1	-0,923	-0,923							-0,473	-0,473							0,046						
0108020110195-1	-0,185	-0,185							-0,767	-0,767							0,051						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108020110197-1	0,976	0,976							1,565	1,565							0,017						
0108020110199-1	-0,031	-0,031							-0,553	-0,553							0,045						
0108020110201-1	0,277	0,277							1,273	1,273							0,023						
0108020110202-1	1,016	1,016							-0,197	-0,197							0,047						
0108020110203-1	-0,344	-0,344							-1,098	-1,098							0,045						
0108020110204-1	-0,307	-0,307							-2,041	-2,041							0,054						
0108020110205-1	0,049	0,049							0,936	0,936							0,035						
0108020110206-1	0,400	0,400							1,032	1,032							0,033						
0108020110208-1	-0,441	-0,441							-1,044	-1,044							0,046						
0108020110209-1	-1,450	-1,450							-0,623	-0,623							0,048						
0108020110211-1	-0,049	-0,049							-0,046	-0,046							0,042						
0108020110213-1	-0,768	-0,768							-1,206	-1,206							0,047						
0108020110215-1	-0,540	-0,540							-0,809	-0,809							0,052						
0108020110217-1	0,215	0,215							0,787	0,787							0,032						
0108020110219-1	0,044	0,044							0,126	0,126							0,041						
0108020110221-1	0,129	0,129							-1,509	-1,509							0,051						
0108020110223	-0,795	-0,795							-1,766	-1,766							0,054						
0108020110226	-0,204	-0,204							-2,188	-2,188							0,053						
0108020110228	-0,091	-0,091							-2,216	-2,216							0,053						
0108020110231	2,775	2,775							4,488	4,488							0,018						
0108020110232	-0,514	-0,514							0,434	0,434							0,039						
0108020110234	-0,342	-0,342							0,596	0,596							0,031						
0108020110235	-0,824	-0,824							-2,086	-2,086							0,054						
0108020110238	0,274	0,274							-0,820	-0,820							0,053						
0108020110239	0,535	0,535							0,079	0,079							0,041						
0108020110243	-0,697	-0,697							-2,932	-2,932							0,054						
0108020110245	-0,740	-0,740							0,050	0,050							0,043						
0108020110246	0,425	0,425							-0,077	-0,077							0,043						
0108020110248	0,276	0,276							-0,228	-0,228							0,045						
0108020110252	-0,826	-0,826							-1,894	-1,894							0,052						
0108020110253	-0,073	-0,073							-0,454	-0,454							0,045						
0108020110254	-0,261	-0,261							-1,409	-1,409							0,050						
0108020110259	-0,078	-0,078							-0,190	-0,190							0,045						
0108020110261	-0,037	-0,037							-0,142	-0,142							0,043						
0108020110262-1	-0,689	-0,689							-0,557	-0,557							0,045						
0108020110263	0,641	0,641							1,155	1,155							0,028						
0108020110266	0,583	0,583							2,217	2,217							0,019						
0108020110267	0,477	0,477							1,130	1,130							0,029						
0108020110268	2,276	2,276							3,309	3,309							0,013						
0108020111002	0,749	0,749							1,619	1,619							0,024						
0108020111003	-0,065	-0,065							0,361	0,361							0,042						
0108020111004	-0,047	-0,047							1,205	1,205							0,027						
0108020111006	1,386	1,386							2,923	2,923							0,013						
0108020111011	-0,068	-0,068							0,945	0,945							0,035						
0108020111012	-0,245	-0,245							-0,911	-0,911							0,050						
0108020111013	0,541	0,541							1,672	1,672							0,024						
0108020111015	-0,106	-0,106							0,346	0,346							0,042						
0108020111016	0,461	0,461							2,018	2,018							0,014						
0108020111017	0,947	0,947							1,666	1,666							0,024						
0108020111018	1,055	1,055							2,069	2,069							0,013						
0108020111019	1,367	1,367							2,622	2,622							0,014						
0108020111022	1,530	1,530							2,484	2,484							0,012						
0108020111023	0,185	0,185							0,899	0,899							0,039						
0108020111024	0,420	0,420							1,819	1,819							0,020						
0108020111026	0,869	0,869							2,304	2,304							0,015						
0108020111028	-0,328	-0,328							0,354	0,354							0,042						
0108020111031	0,059	0,059							0,878	0,878							0,038						
0108020111032	0,293	0,293							1,885	1,885							0,020						
0108020111033	0,146	0,146							0,945	0,945							0,035						
0108020111034	0,033	0,033							0,652	0,652							0,030						
0108020115013	0,045	0,045							0,285	0,285							0,040						
0108020115026	-0,444	-0,444							-1,690	-1,690							0,054						
0108020115027	0,274	0,274							1,057	1,057							0,031						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108020115028	0,218	0,218							1,113	1,113							0,030						
0108020115029	-1,214	-1,214							-3,326	-3,326							0,054						
0108020115032	-0,665	-0,665							-2,783	-2,783							0,052						
0108020115033	-2,057	-2,057							-2,094	-2,094							0,054						
0108020115034	-0,764	-0,764							-1,621	-1,621							0,053						
0108020115035	-0,657	-0,657							-0,011	-0,011							0,043						
0108020115039	-0,405	-0,405							-2,296	-2,296							0,053						
0108020115040	-0,339	-0,339							-3,425	-3,425							0,055						
0108020115042	-1,043	-1,043							-2,035	-2,035							0,054						
0108020115045	-1,390	-1,390							-1,567	-1,567							0,052						
0108020115049	-0,153	-0,153							-0,971	-0,971							0,050						
0108020115051	-1,028	-1,028							-2,499	-2,499							0,053						
0108020115054	-0,863	-0,863							-2,847	-2,847							0,054						
0108020115056	-1,319	-1,319							-2,000	-2,000							0,054						
0108020115060	-0,737	-0,737							-2,804	-2,804							0,053						
0108020115063	-1,006	-1,006							-2,976	-2,976							0,054						
0108020115065	-1,393	-1,393							-1,919	-1,919							0,052						
0108020115066	-1,978	-1,978							-0,302	-0,302							0,049						
0108020115067	-1,925	-1,925							-2,054	-2,054							0,054						
0108020115072	1,385	1,385							1,947	1,947							0,013						
0108020115076	-0,551	-0,551							-0,759	-0,759							0,050						
0108020115086	-0,093	-0,093							0,035	0,035							-8,397						
010803000301235572-2	0,695	0,695							0,864	0,864							0,017						
010803000301235573-1	1,010	1,010							1,388	1,388							0,015						
010803000301235578-1	0,919	0,919							1,376	1,376							0,015						
010803000301235578-2	3,701	3,701							3,657	3,657							0,017						
010803000301235579-1	0,614	0,614							1,183	1,183							0,015						
010803000301235582-1	0,344	0,344							1,072	1,072							0,015						
010803000301235582-2	2,665	2,665							3,421	3,421							0,015						
010803000301235584-1	-0,858	-0,858							-0,769	-0,769							0,019						
010803000301235585-2	2,845	2,845							3,436	3,436							0,015						
010803000301235586-2	3,986	3,986							4,330	4,330							0,018						
010803000301235587-1	2,097	2,097							2,541	2,541							0,013						
010803000301235587-2	4,092	4,092							4,816	4,816							0,019						
010803000301235592-2	3,188	3,188							3,807	3,807							0,017						
010803000301235594-1	-0,364	-0,364							0,214	0,214							0,018						
010803000301235602-1	1,137	1,137							1,413	1,413							0,015						
010803000301235602-2	3,183	3,183							3,479	3,479							0,016						
010803000301235603-2	3,746	3,746							4,502	4,502							0,019						
010803000301235604-2	1,533	1,533							1,783	1,783							0,012						
010803000301235605-2	2,856	2,856							3,738	3,738							0,017						
010803000301235606-1	1,013	1,013							1,383	1,383							0,015						
010803000301235606-2	3,254	3,254							4,397	4,397							0,018						
010803000301235607-1	0,427	0,427							0,758	0,758							0,016						
010803000301235607-2	1,102	1,102							2,024	2,024							0,012						
010803000301235608-1	0,596	0,596							0,768	0,768							0,016						
010803000301235608-2	2,329	2,329							2,654	2,654							0,014						
010803000301235676-1	-0,367	-0,367							0,009	0,009							0,019						
010803000301235738-1	-0,711	-0,711							-0,470	-0,470							0,019						
010803000301235739-1	0,276	0,276							0,902	0,902							0,016						
010803000301235740-1	1,529	1,529							1,764	1,764							0,016						
010803000301235740-2	1,211	1,211							1,918	1,918							0,013						
010803000301235768-1	1,844	1,844							2,800	2,800							0,013						
010803000301235768-2	2,935	2,935							4,191	4,191							0,019						
010803000301235812-1	4,525	4,525							5,503	5,503							0,020						
010803000301235812-2	5,241	5,241							4,700	4,700							0,020						
010803000301235813-2	-0,564	-0,564							-0,319	-0,319							0,019						
010803000301235905-1	0,933	0,933							1,884	1,884							0,013						
010803000301236033-2	0,433	0,433							0,860	0,860							0,017						
010803000301236034-1	0,144	0,144							0,411	0,411							0,015						
010803000301236036-1	1,519	1,519							1,556	1,556							0,014						
010803000301236038-1	3,332	3,332							3,578	3,578							0,017						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
010803000301237450-3	0,601	0,601							0,601	0,601							0,017						
010803000301238301-1	-0,336	-0,336							0,014	0,014							0,019						
010803000301238303-1	-0,205	-0,205							0,585	0,585							0,015						
010803000301238307-1	0,586	0,586							1,153	1,153							0,015						
010803000301238346-1	-1,117	-1,117							-1,177	-1,177							0,019						
010803000301238347-1	0,240	0,240							0,130	0,130							0,019						
010803000301238348-1	-1,239	-1,239							-0,549	-0,549							0,019						
010803000301238349-1	1,719	1,719							1,621	1,621							0,013						
010803000301238350-1	1,305	1,305							0,853	0,853							0,018						
010803000301238567-1	0,484	0,484							0,342	0,342							0,017						
010803000301238568-1	-0,015	0,153	-2,021	-3,127	-4,058	-3,448	-3,092	-0,107	0,096	-0,406	-0,818	-0,728	-0,289	0,342	1,010	1,56	0,196	0,381	0,540	0,687	0,861	1,098	0,125
010803000301238628-1	1,524	1,524							1,029	1,029							0,015						
010803000301238640-1	-0,025	-1,486	-1,776	-2,380	-1,201	-0,126			0,194	-1,120	-0,715	0,132	1,041	1,631			0,121	0,213	0,304	0,439	0,084		
010803000301238646-1	-0,509	-0,755	-3,136	-3,494	-3,854	-2,544			-0,116	-0,872	-0,891	-0,391	0,384	1,192			0,192	0,368	0,521	0,691	0,094		
010803000301238707-1	-0,063	-0,957	-2,317	-2,435	-2,309	-0,316			0,062	-0,660	-0,677	-0,089	0,651	1,086			0,141	0,255	0,359	0,480	0,087		
010803000301238708-1	1,280	1,280							0,676	0,676							0,016						
010803000301238728-1	-0,073	-0,609	-1,975	-1,773	-1,641	-0,364			0,164	-0,457	-0,369	-0,101	0,433	1,316			0,118	0,221	0,321	0,442	0,084		
010803000301238729-1	-0,279	0,532	-2,139	-3,007	-4,199	-3,938	-4,199	-1,953	0,009	-0,286	-0,727	-0,712	-0,373	0,157	0,745	1,26	0,218	0,428	0,602	0,761	0,925	1,118	0,126
010803000301238768-1	-0,388	-0,789	-2,529	-2,979	-3,060	-1,939			-0,089	-0,660	-0,792	-0,143	0,560	0,591			0,128	0,242	0,343	0,456	0,086		
010803000301238769-1	0,613	0,613							0,100	0,100							0,019						
010803000301238838-1	2,140	2,140							1,395	1,395							0,015						
010803000301238840-1	1,595	1,595							0,851	0,851							0,018						
010803000301238841-1	0,890	0,890							0,584	0,584							0,015						
010803000301238844-1	-0,079	-1,522	-1,971	-2,586	-1,799	-0,393			0,164	-0,927	-0,423	0,058	0,645	1,467			0,130	0,229	0,323	0,443	0,085		
010803000301238854-1	-0,443	-1,032	-3,000	-4,301	-3,612	-2,214			0,041	-1,065	-1,408	-0,189	1,260	1,605			0,152	0,290	0,414	0,585	0,091		
010803000301238855-1	0,841	0,841							0,272	0,272							0,018						
010803000301238868-1	1,448	1,448							0,931	0,931							0,016						
010803000301238871-1	-0,251	-1,102	-2,798	-2,972	-3,212	-1,253			-0,025	-0,537	-0,754	-0,398	0,331	1,233			0,204	0,376	0,530	0,704	0,094		
010803000301238874-1	-0,582	-0,614	-2,929	-3,066	-3,805	-2,909			-0,182	-0,313	-0,625	-0,641	-0,198	0,867			0,174	0,338	0,489	0,653	0,093		
010803000301238875-1	-0,250	0,220	-2,138	-2,885	-4,015	-3,657	-2,724	-1,749	0,091	-0,590	-0,870	-0,646	-0,118	0,515	1,053	1,30	0,214	0,418	0,591	0,752	0,934	1,215	0,128
010803000301238988-1	1,557	1,557							0,805	0,805							0,016						
010803000301238989-1	0,221	0,221							0,284	0,284							0,018						
010803000301238990-1	-0,324	0,281	-2,532	-3,472	-3,502	-1,618			-0,160	-0,291	-1,023	-0,804	0,058	1,259			0,138	0,270	0,381	0,504	0,088		
010803000301238992-1	-0,240	0,228	-1,971	-3,104	-4,218	-4,314	-3,840	-1,679	-0,043	0,083	-0,506	-0,766	-0,697	-0,301	0,420	1,47	0,201	0,395	0,566	0,723	0,883	1,095	0,126
010803000301239051-1	-0,261	-0,876	-1,897	-2,854	-1,579	-1,307			0,085	-0,571	-0,900	-0,120	0,865	1,153			0,148	0,279	0,403	0,597	0,090		
010803000301239061-1	-0,155	-0,485	-2,591	-3,113	-3,514	-0,777			-0,013	-0,245	-0,872	-0,933	-0,074	2,058			0,147	0,280	0,396	0,520	0,089		
010803000301239062-1	-0,276	0,275	-2,490	-2,689	-2,606	-1,381			-0,258	-0,326	-0,930	-0,546	0,127	0,385			0,177	0,346	0,499	0,687	0,093		
010803000301239063-1	0,356	-1,963	-1,983	-2,320	-0,367	1,779			0,207	-1,676	-0,687	0,239	1,133	2,026			0,115	0,185	0,251	0,357	0,077		
010803000301241375-1	-0,334	0,268	-1,569	-1,667	-1,981	-1,670			0,155	-0,008	-0,109	0,076	0,340	0,477			0,102	0,201	0,298	0,407	0,083		
010803000301241378-1	-0,405	-1,074	-1,658	-2,694	-2,899	-2,024			0,187	-0,825	-0,346	0,326	0,859	0,920			0,064	0,124	0,181	0,239	0,069		
010803000301241381-1	-0,414	-1,694	-2,412	-3,356	-2,961	-2,068			0,054	-0,570	-0,446	0,073	0,573	0,642			0,124	0,228	0,324	0,434	0,086		
010803000301241383-1	-0,290	-1,607	-1,920	-2,584	-2,583	-1,449			0,242	-0,615	0,371	0,473	0,344	0,638			0,063	0,118	0,170	0,225	0,064		
010803000301241389-1	-0,259	-0,984	-2,328	-3,306	-2,956	-1,295			0,010	-0,634	-0,907	-0,207	0,710	1,089			0,183	0,339	0,478	0,641	0,092		
010803000301241390-1	-0,266	-0,266							0,573	0,573							0,015						
010803000301241391-1	-0,008	-0,008							-0,550	-0,550							0,019						
010803000301241409-1	-0,246	-0,578	-2,659	-2,844	-3,064	-1,231			0,259	-0,778	-0,205	0,197	0,661	1,422			0,129	0,245	0,338	0,432	0,086		
010803000301241410-1	-0,332	-0,356	-2,501	-2,399	-2,557	-1,661			-0,047	-0,662	-0,586	-0,099	0,440	0,672			0,120	0,231	0,330	0,445	0,086		
010803000301241412-1	-0,803	-0,734	-2,002	-3,530	-3,519	-4,017			-0,220	0,356	-0,800	-0,402	0,178	-0,431			0,118	0,233	0,346	0,481	0,088		
010803000301241413-1	-0,221	-0,842	-2,713	-2,611	-2,955	-1,104			0,218	-0,717	-0,231	0,144	0,594	1,300			0,132	0,247	0,341	0,439	0,086		
010803000301241414-1	-0,548	-0,280	-2,312	-2,705	-3,155	-2,742			-0,147	-0,002	-0,359	-0,136	0,075	-0,313			0,111	0,216	0,315	0,424	0,085		
010803000301241415-1	-0,341	-1,371	-1,986	-3,242	-2,376	-1,707			0,107	-0,612	-0,430	0,110	0,646	0,819			0,122	0,227	0,324	0,442	0,086		
010803000301241593-1	-0,464	-0,969	-2,487	-3,640	-3,029	-2,319			-0,065	-0,569	-0,743	-0,195	0,482	0,699			0,127	0,240	0,340	0,462	0,087		
010803000301241597-1	-0,167	-0,996	-2,230	-2,217	-2,251	-0,837			0,213	-0,616	-0,463	0,112	0,788	1,244			0,125	0,230	0,326	0,434	0,085		
010803000301241806-1	-0,295	-1,460	-2,163	-2,011	-1,180				0,025	-0,953	-0,270	0,313	1,010				0,106	0,198	0,300	0,070			
010803000301241807-1	-0,171	-1,515	-1,808	-2,707	-2,550	-0,855			0,136	-0,652	0,232	0,130	0,034	0,935			0,132	0,238	0,336	0,440	0,085		
010803000301241808-1	-0,315	-0,487	-1,849	-2,074	-2,255	-1,573			-0,004	-0,563	-0,139	-0,025	0,110	0,597			0,098	0,189	0,277	0,376	0,080		
010803000301241811-1	-0,634	-0,475	-1,659	-2,375	-2,927	-3,168			0,034	0,053	0,186	-0,221	-0,373	0,523			0,094	0,185	0,277	0,379	0,081		
010803000301241812-1	0,077	-1,845	-2,350	-3,518	-2,694	0,388			0,124	-1,310	-0,854	-0,082	0,896	1,971			0,233	0,400	0,554	0,749	0,094		
010803000301241917-1	-0,365	-0,592	-1,737	-2,360	-2,326	-1,827			0,132	-0,033	-0,278	0,052	0,462	0,459			0,105	0,203	0,297	0,404	0,083		
010803000301241921-1	-0,451	-0,717	-2,357	-3,387	-2,725	-2,256			-0,019	-0,247	-0,602	-0,350	0,233	0,872			0,132	0,252	0,360	0,498	0,088		
010803000301241924-1	-0,371	-0,596	-2,245	-2,222	-2,314	-1,854			0,008	-0,469	-0,495	0,054	0,558	0,395			0,108	0,208	0,304	0,415	0,084		
010803000301241925-1	-0,558	-0,512	-1,983	-2,499	-3,538	-2,789			-0,176	0,477	-0,534	-0,857	-0,499	0,533			0,162	0,317	0,468	0,626	0,092		
010803000301241926-1	0,272	-1,086	-1,464	-1,158	-0,250	1,359			0,324	-0,945	0,172	0,707	0,859	0,826			0,067	0,122	0,173	0,235	0,063		

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
010803000301242169-1	-0,142	-0,629	-1,874	-2,564	-2,622	-2,074	-0,852		0,044	-0,367	-0,700	-0,404	0,179	0,710	0,847		0,129	0,241	0,342	0,447	0,588	0,099	
010803000301242171-1	-0,627	-1,215	-2,593	-3,866	-3,659	-3,133			-0,079	-0,308	-0,637	-0,174	0,384	0,338			0,126	0,241	0,344	0,461	0,088		
010803000301242180-1	0,103	-0,658	-1,521	-2,354	-2,844	0,515			0,358	0,601	-0,514	-1,287	-0,415	3,404			0,166	0,318	0,465	0,617	0,091		
010803000301242204-1	-0,196	-0,845	-2,636	-2,647	-2,364	-0,979			-0,006	-0,880	-0,676	0,097	0,767	0,663			0,131	0,244	0,339	0,449	0,086		
010803000301242207-1	-0,687	-0,431	-1,144	-1,569	-2,962	-3,434			-0,237	1,151	0,057	-0,524	-0,818	-1,053			0,109	0,219	0,333	0,455	0,084		
010803000301242228-1	-0,410	-0,477	-3,306	-3,662	-3,758	-2,049			-0,015	-0,980	-0,932	-0,150	0,767	1,221			0,200	0,382	0,521	0,675	0,093		
010803000301242234-1	-0,547	-1,374	-2,124	-3,298	-2,985	-2,735			-0,147	0,042	-0,599	-0,541	-0,090	0,453			0,149	0,284	0,415	0,576	0,091		
010803000301242250-1	-0,131	-0,673	-2,158	-2,186	-2,649	-0,657			0,258	-0,382	-0,241	-0,150	0,349	1,713			0,128	0,239	0,336	0,440	0,085		
010803000301242254-1	0,157	-0,680	-1,297	-1,375	-1,245	-0,472	0,942		0,178	-0,427	-0,063	0,132	0,261	0,428	0,735		0,114	0,212	0,303	0,400	0,526	0,094	
010803000301242264-1	0,005	-1,096	-1,779	-1,873	-1,476	0,024			0,143	-0,349	-0,189	0,228	0,560	0,465			0,113	0,205	0,293	0,395	0,082		
010803000301242269-1	-0,197	-0,802	-1,641	-2,561	-3,131	-2,872	-1,184		-0,049	0,166	-0,406	-0,690	-0,584	0,014	1,204		0,186	0,354	0,514	0,674	0,872	0,109	
010803000301242272-1	0,036	-0,341	-2,870	-2,888	-2,488	0,183			0,083	-1,351	-1,000	-0,009	1,075	1,702			0,213	0,403	0,560	0,761	0,094		
010803000301242274-1	-0,851	-0,816	-1,995	-3,884	-3,618	-4,257			-0,235	0,390	-0,599	-0,880	-0,516	0,430			0,146	0,290	0,431	0,612	0,092		
010803000301242310-1	-0,477	-0,124	-2,479	-2,828	-3,631	-2,385			-0,021	-0,019	-0,507	-0,493	-0,007	0,923			0,147	0,286	0,408	0,534	0,090		
010803000301242313-1	-0,444	-1,079	-1,896	-3,190	-3,362	-2,222			-0,211	0,306	-0,340	-0,910	-0,781	0,668			0,166	0,319	0,467	0,632	0,092		
010803000301242315-1	-0,252	-0,910	-2,339	-2,992	-3,497	-3,069	-1,513		0,044	0,022	-0,396	-0,517	-0,310	0,255	1,208		0,169	0,314	0,441	0,566	0,722	0,106	
010803000301242596-1	-0,139	-1,300	-2,178	-2,619	-2,424	-0,696			-0,025	-0,230	-0,559	-0,570	-0,035	1,269			0,188	0,344	0,496	0,682	0,093		
010803000301242597-1	-0,778	-0,486	-1,775	-2,268	-3,746	-3,890			-0,250	1,120	0,119	-0,710	-1,078	-0,699			0,126	0,250	0,375	0,506	0,089		
010803000301242601-1	-0,571	-1,712	-2,612	-3,792	-3,986	-2,855			-0,101	0,113	-0,690	-0,659	-0,070	0,802			0,189	0,351	0,502	0,667	0,093		
010803000301242603-1	-0,164	-1,171	-1,699	-2,057	-1,704	-0,818			0,070	-0,448	-0,399	0,099	0,568	0,530			0,114	0,213	0,308	0,423	0,084		
010803000301242744-1	-0,797	-0,754	-2,110	-3,666	-3,838	-3,985			-0,271	0,619	-0,432	-0,649	-0,489	-0,405			0,141	0,278	0,410	0,561	0,091		
010803000301243169-1	-0,406	-0,268	-1,647	-3,369	-2,820	-2,030			0,016	-0,440	-0,826	-0,482	0,360	1,471			0,172	0,335	0,488	0,673	0,093		
010803030060910301-3	-0,928	-0,531	-1,883	-2,985	-3,713				-0,814	-0,458	-0,654	-1,372	-0,772				0,106	0,216	0,340	0,070			
010803030060910317	-0,345	-0,345							-0,288	-0,288							0,019						
010803030060910322	0,210	0,210							0,051	0,051							0,019						
010803030060910323	-0,444	-0,444							-0,417	-0,417							0,019						
010803030060910325	-0,836	-0,821	-1,636	-2,509					-0,910	-1,316	-0,974	-0,439					0,111	0,244	0,055				
010803030060910327	-0,660	-0,927	-1,872	-2,617	-2,641				-0,549	-0,958	-0,841	-0,607	0,211				0,103	0,207	0,322	0,070			
010803030060940009-1	0,494	-1,174	-0,333	1,483					0,235	-1,250	0,630	1,323					0,058	0,113	0,045				
010803030060940011-1	-1,028	-1,028							-1,603	-1,603							0,019						
0108030310005-7	0,865	0,865							-0,757	-0,757							0,019						
0108030310006-3	0,100	0,100							0,898	0,898							0,017						
0108030310009-2	0,683	0,683							1,199	1,199							0,015						
0108030310011-1	-0,616	-1,513	-2,121	-3,453	-2,465				-0,285	-1,083	-0,179	-0,889	1,011				0,149	0,287	0,427	0,075			
0108030310015-2	-0,567	-1,573	-2,289	-1,699					-0,071	-0,399	-0,530	0,715					0,115	0,225	0,056				
0108030310016-2	-0,654	-1,139	-2,122	-1,962					-0,358	-0,730	-0,455	0,112					0,144	0,291	0,057				
0108030310017-2	-0,790	-1,345	-2,392	-2,369					0,025	0,154	-0,543	0,465					0,098	0,197	0,055				
0108030310018-2	0,499	0,499							0,688	0,688							0,016						
0108030310020-2	-0,039	-0,039							-0,857	-0,857							0,019						
0108030310024-3	-1,548	-1,548							-1,728	-1,728							0,018						
0108030310372	0,821	0,821							1,561	1,561							0,013						
0108030310373	-1,048	-0,886	-1,967	-3,262	-4,191				-2,062	-2,277	-1,662	-2,047	-2,261				0,075	0,156	0,252	0,062			
0108030310606-1	-0,539	-0,539							0,507	0,507							0,015						
0108030310609-1	1,015	1,015							-0,796	-0,796							0,020						
0108030310611-2	-0,759	-0,730	-1,604	-2,812	-3,803	-3,795			-0,343	0,106	-1,004	-0,994	-0,333	0,509			0,104	0,208	0,312	0,425	0,085		
0108030310612-1	-0,480	-0,480							-0,783	-0,783							0,019						
0108030310613-2	-0,209	-1,230	-2,088	-2,505	-0,836				-0,304	-1,195	-1,364	0,008	1,334				0,114	0,223	0,347	0,072			
0108030311006	-0,859	-0,859							-1,600	-1,600							0,019						
0108030311017	-0,148	-0,148							-1,232	-1,232							0,019						
0108030311018	0,190	0,190							0,182	0,182							0,018						
0108030311020	-0,060	-0,060							-1,133	-1,133							0,019						
0108030311028	-0,989	-0,989							-0,911	-0,911							0,019						
0108030311030	0,010	-1,214	-1,284	0,029					0,165	-0,617	0,269	0,843					0,086	0,166	0,052				
0108030320002	-0,520	-1,210	-1,620	-1,561					-0,045	-0,420	-0,317	0,600					0,101	0,207	0,054				
0108030320015	0,289	0,289							0,103	0,103							0,019						
0108030320016	-0,494	-1,100	-1,836	-1,482					-0,769	-1,229	-0,980	-0,099					0,108	0,226	0,055				
0108030320019-2	-0,060	-0,850	-1,245	-0,181					-0,164	-1,130	-0,290	0,929					0,118	0,238	0,056				
0108030320020-2	0,545	-0,894	-0,793	0,037	2,178				1,023	-0,422	0,584	1,380	2,552				0,055	0,101	0,150	0,051			
0108030320021-1	-0,618	-0,618							-0,534	-0,534							0,019						
0108030320022-1	-0,796	-0,796							-1,930	-1,930							0,020						
0108030320022-3	-1,441	-1,441							-2,715	-2,715							0,019						
0108030320023-2	0,570	0,570							0,286	0,286							0,018						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108030320025	-0,451	-0,451							-0,624	-0,624							0,019						
0108030320026-2	-0,367	-0,367							-0,618	-0,618							0,019						
0108030320029	-2,372	-2,372							-4,356	-4,356							0,019						
0108030320030	-1,463	-1,463							-1,920	-1,920							0,019						
0108030320031	-1,625	-1,625							-1,606	-1,606							0,019						
0108030320033	-1,539	-1,539							-2,102	-2,102							0,019						
0108030320035	-1,067	-1,067							-0,333	-0,333							0,019						
0108030320036	-0,561	-0,561							-1,516	-1,516							0,019						
0108030320038	-1,202	-1,202							-1,311	-1,311							0,019						
0108030320039	0,327	0,327							0,539	0,539							0,015						
0108030320040	-0,907	-0,907							-2,541	-2,541							0,019						
0108030320041	-0,815	-0,815							-0,738	-0,738							0,019						
0108030320042	-1,110	-1,110							-1,164	-1,164							0,019						
0108030320045	-0,806	-0,806							-1,444	-1,444							0,019						
0108030320047-1	-0,655	-0,655							-0,530	-0,530							0,019						
0108030320047-2	0,270	0,270							0,544	0,544							0,015						
0108030320049-2	1,002	1,002							0,921	0,921							0,016						
0108030320049-3	1,304	1,304							0,609	0,609							0,018						
0108030320050-1	-0,268	-0,268							-0,378	-0,378							0,019						
0108030320050-2	-0,933	-0,933							-0,929	-0,929							0,019						
0108030320050-3	1,673	1,673							1,259	1,259							0,015						
0108030320051-2	-0,493	-0,493							-0,970	-0,970							0,019						
0108030320051-3	-1,019	-1,019							-1,178	-1,178							0,019						
0108030320052-1	0,419	0,419							0,408	0,408							0,015						
0108030320052-3	-0,115	-0,115							0,178	0,178							0,018						
0108030320053-1	0,839	0,839							0,904	0,904							0,016						
0108030320053-2	0,578	-1,035	-0,767	1,734					0,424	-1,015	0,112	2,173					0,112	0,224	0,056				
0108030320060-1	-0,679	-0,679							-1,546	-1,546							0,019						
0108030320061-1	-0,578	-0,578							-0,937	-0,937							0,019						
0108030320061-2	-0,328	-0,328							-0,064	-0,064							0,018						
0108030320062-1	0,386	0,386							0,793	0,793							0,016						
0108030320062-2	-0,662	-0,662							0,186	0,186							0,018						
0108030320062-3	-0,171	-0,171							-0,514	-0,514							0,019						
0108030320063-1	-0,263	-0,263							-0,259	-0,259							0,019						
0108030320066-3	-0,510	-0,763	-1,894	-1,530					-0,894	-0,863	-1,802	-0,016					0,076	0,157	0,051				
0108030320069-1	-0,392	-0,392							-1,558	-1,558							0,020						
0108030320069-2	-0,931	-0,931							-1,735	-1,735							0,018						
0108030320069-3	1,206	1,206							1,183	1,183							0,015						
0108030320070-2	-0,308	-0,308							0,097	0,097							0,019						
0108030320071-1	-0,456	-0,456							-0,650	-0,650							0,019						
0108030320071-2	-0,548	-0,548							-0,346	-0,346							0,019						
0108030320072-1	0,654	0,654							0,738	0,738							0,016						
0108030320073-1	-0,685	-0,685							-0,455	-0,455							0,019						
0108030606132 52-4	-0,262	-0,262							-2,377	-2,377							0,018						
0108030606132 62-1	-0,634	-0,634							-0,340	-0,340							0,019						
0108030606132 62-4	-0,681	-0,681							0,036	0,036							0,019						
0108030606132 63-1	-1,252	-1,252							-1,955	-1,955							0,019						
01080306061330007	-0,395	-0,395							-0,344	-0,344							0,019						
01080306061330019	-0,725	-0,725							-0,363	-0,363							0,019						
01080306061330025-2	-0,203	-0,203							0,654	0,654							0,016						
01080306061330034-2	-1,068	-1,068							-2,022	-2,022							0,019						
01080306061330037	-0,774	-0,774							-0,543	-0,543							0,019						
01080306061330044-3	0,203	0,203							-0,023	-0,023							0,018						
01080306061330047	0,483	0,483							0,614	0,614							0,017						
01080306061330060-2	-0,613	-0,613							-1,527	-1,527							0,019						
01080306061330063-2	-0,389	-0,389							0,201	0,201							0,018						
01080306061330065-1	0,068	-1,071	-1,101	0,203					0,379	-0,888	0,245	1,781					0,095	0,185	0,054				
01080306061330069-2	-0,506	-0,842	-1,651	-1,518					-0,897	-1,341	-1,133	-0,218					0,087	0,182	0,053				
01080306061330075	0,272	0,272							0,510	0,510							0,015						
01080306061330078	-0,014	-0,014							0,730	0,730							0,016						
01080306061330079	-0,353	-0,353							-0,565	-0,565							0,019						
01080306061330083	-0,703	-0,703							-1,488	-1,488							0,019						
01080306061330089	-0,272	-0,272							-0,176	-0,176							0,019						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
01080306061330093	-0,304	-0,304							-0,174	-0,174							0,019						
01080306061330094	0,143	0,143							0,158	0,158							0,018						
01080306061330095	-0,309	-0,309							-0,754	-0,754							0,019						
01080306061330097	-0,084	-0,084							-0,174	-0,174							0,019						
01080306061330099	-0,759	-0,759							-0,805	-0,805							0,019						
01080306061330100	-0,341	-0,341							-0,924	-0,924							0,019						
01080306061330101	-0,534	-0,534							-1,374	-1,374							0,019						
01080306061330104	-1,224	-1,224							-2,107	-2,107							0,019						
01080306061330109	-1,591	-1,591							-2,987	-2,987							0,019						
01080306061330112	-0,551	-0,551							-0,415	-0,415							0,019						
01080306061330113	0,157	0,157							-0,271	-0,271							0,019						
01080306061330114	-0,859	-0,859							-1,112	-1,112							0,019						
01080306061330115	-0,104	-0,104							-0,213	-0,213							0,019						
01080306061330116	-1,042	-1,042							-0,788	-0,788							0,019						
01080306061330121	0,054	0,054							0,261	0,261							0,018						
01080306061330122	-0,330	-0,330							0,090	0,090							0,019						
01080306061330123	-0,140	-0,140							0,330	0,330							0,017						
0108030606133 53-1	-1,011	-1,011							-1,428	-1,428							0,019						
01080306061340005-2	-0,472	-0,472							0,476	0,476							0,015						
01080306061340006-1	0,275	0,275							0,312	0,312							0,018						
01080306061340013-1	-0,637	-0,651	-1,399	-1,912					-0,950	-1,310	-0,839	-0,700					0,115	0,252	0,055				
01080306061340014	-0,180	-0,180							-0,097	-0,097							0,018						
01080306061360 1-3	-2,332	-2,332							-3,409	-3,409							0,019						
0108030606136 04-2	-1,228	-1,228							-1,020	-1,020							0,019						
01080306061390 3-2	-0,420	-0,420							-0,733	-0,733							0,019						
01080306061390 9-1	-0,857	-0,857							-0,886	-0,886							0,019						
0108030610350-1	-0,336	-0,336							0,358	0,358							0,018						
0108030610352	-0,538	-0,538							-1,312	-1,312							0,019						
0108030610354-1	-0,855	-0,855							-0,774	-0,774							0,019						
0108030610355	0,790	0,790							1,345	1,345							0,015						
0108030610358	-0,963	-0,963							-0,235	-0,235							0,019						
0108030610364-1	-0,045	-0,045							0,411	0,411							0,015						
0108030610364-2	-0,086	-0,086							-0,423	-0,423							0,019						
0108030610368-1	-1,244	-1,244							-1,562	-1,562							0,019						
0108030610370	1,083	1,083							-0,394	-0,394							0,019						
0108030610383	0,241	0,241							0,942	0,942							0,016						
0108030610396	-1,127	-1,127							-1,150	-1,150							0,019						
0108030610700-1	0,075	0,075							-0,910	-0,910							0,019						
0108030610700-2	-0,614	-0,614							-0,333	-0,333							0,019						
0108030610700-3	0,155	-1,287	-2,478	-1,962	0,620				-1,082	-1,923	-2,212	-0,805	0,613				0,078	0,151	0,245	0,065			
0108030610702-2	-0,288	-0,288							-0,067	-0,067							0,018						
0108030610703-2	-0,462	-0,462							-0,416	-0,416							0,019						
0108030610705-1	-1,410	-1,410							-1,366	-1,366							0,019						
0108030610705-2	-0,425	-0,425							0,126	0,126							0,019						
0108030610706-3	-0,751	-0,809	-1,570	-2,801	-3,002				-0,246	1,596	-1,218	-1,111	-0,252				0,145	0,290	0,447	0,075			
0108030612006-1	-1,461	-1,461							-2,179	-2,179							0,018						
0108030612006-2	-1,097	-0,510	-1,525	-2,239	-3,746	-5,483			-1,127	-1,399	-0,520	-0,733	-1,340	-1,643			0,102	0,207	0,322	0,452	0,081		
0108030612007-1	-0,393	-0,393							0,247	0,247							0,018						
0108030612011-1	1,766	1,766							1,509	1,509							0,014						
0108030612012-1	-0,626	-0,626							-1,538	-1,538							0,019						
0108030612013-3	-0,264	-0,264							-0,482	-0,482							0,019						
0108030612014-1	0,589	0,589							0,782	0,782							0,016						
0108030612014-2	-0,485	-0,810	-0,917	-1,184	-1,941				-0,626	-1,108	-0,659	0,612	-1,348				0,148	0,310	0,513	0,074			
0108030612014-3	-0,155	-0,573	-1,415	-1,879	-1,521	-0,775			-0,287	-0,725	-0,737	-0,344	0,106	0,266			0,144	0,277	0,411	0,580	0,090		
0108030612015	-0,552	-0,552							-0,750	-0,750							0,019						
0108030612019-1	-0,352	-0,352							-1,937	-1,937							0,019						
0108030612019-2	-0,050	-0,050							-0,102	-0,102							0,019						
0108030612019-3	-0,905	-1,479	-2,767	-1,696	-3,621				-0,305	-0,733	-1,396	2,318	-1,410				0,078	0,157	0,280	0,064			
0108030612024-2	-0,571	-0,571							-1,463	-1,463							0,019						
0108030612027	-0,493	-0,493							-0,509	-0,509							0,019						
0108030612038	0,135	0,135							-0,118	-0,118							0,019						
0108030612041	-0,046	-0,046							0,672	0,672							0,016						

Tabel C.1: Item-parametre *(continued)*

Opgavenummer	Beta 2017	t.1	t.2	t.3	t.4	t.5	t.6	t.7	Beta dnt	dnt.1	dnt.2	dnt.3	dnt.4	dnt.5	dnt.6	dnt.7	se.t.1	se.t.2	se.t.3	se.t.4	se.t.5	se.t.6	se.t.7
0108030612047	-1,178	-1,178							0,051	0,051							0,019						
0108030615015-1	-0,769	-0,769							-0,805	-0,805							0,019						
0108030615015-2	-0,247	-0,247							0,213	0,213							0,018						
0108030615016-1	-0,953	-0,953							-0,665	-0,665							0,019						
0108030615016-2	-0,428	-0,428							-0,284	-0,284							0,019						
01080306904-1__2	0,333	0,333							0,572	0,572							0,015						
01080306910-1__2	1,260	1,260							1,054	1,054							0,014						
01080306912-1__2	-0,783	-0,783							-2,330	-2,330							0,018						
01080306913-1__2	0,327	0,327							0,073	0,073							0,019						
01080306916-1__2	-1,114	-1,114							-0,624	-0,624							0,019						
01080306920-1__2	1,352	1,352							1,499	1,499							-5,297						

D

Percentiler i den nye analyse fra 2017 og DNT 2017

I denne tabel er elevernes resultat fordelt i 100 lige store grupper af elever (percentilgrupper) med dygtighed i samme interval, og det er angivet hvad estimatet for den dygtigste elev i gruppen er. Percentilerne for DNT er baseret på de aktuelle dygtigheder i 2017, ikke de historiske dygtigheder som rapporteres til lærerne i nationale tests system (og som er en måde at muliggøre sammenligning over tid på).

Tabel D.1: Percentiler

	Profilområde 1		Profilområde 2		Profilområde 3	
Percentil	2017	DNT	2017	DNT	2017	DNT
1	-2,69	-2,89	-1,11	-1,67	-1,77	-2,59
2	-2,23	-2,26	-0,76	-0,85	-1,40	-2,00
3	-1,99	-1,91	-0,57	-0,38	-1,18	-1,59
4	-1,82	-1,69	-0,43	-0,13	-1,02	-1,26
5	-1,69	-1,51	-0,32	0,06	-0,91	-0,96
6	-1,57	-1,35	-0,23	0,22	-0,82	-0,75
7	-1,47	-1,24	-0,15	0,37	-0,73	-0,59
8	-1,40	-1,16	-0,08	0,48	-0,67	-0,47
9	-1,33	-1,09	-0,02	0,57	-0,61	-0,37
10	-1,26	-1,04	0,05	0,65	-0,55	-0,29
11	-1,20	-0,98	0,11	0,74	-0,51	-0,23
12	-1,15	-0,93	0,17	0,83	-0,46	-0,17
13	-1,10	-0,89	0,22	0,92	-0,41	-0,12
14	-1,05	-0,84	0,27	0,99	-0,37	-0,07
15	-1,01	-0,80	0,32	1,05	-0,33	-0,02
16	-0,96	-0,76	0,37	1,12	-0,29	0,02
17	-0,92	-0,73	0,41	1,18	-0,25	0,06
18	-0,88	-0,69	0,46	1,23	-0,22	0,09
19	-0,84	-0,66	0,49	1,28	-0,18	0,13
20	-0,80	-0,62	0,53	1,32	-0,15	0,16
21	-0,76	-0,59	0,57	1,37	-0,12	0,19
22	-0,73	-0,56	0,60	1,41	-0,08	0,23
23	-0,70	-0,53	0,64	1,45	-0,05	0,26
24	-0,66	-0,50	0,67	1,49	-0,02	0,29
25	-0,63	-0,47	0,71	1,54	0,01	0,32
26	-0,60	-0,44	0,74	1,59	0,04	0,34
27	-0,57	-0,42	0,77	1,63	0,07	0,37
28	-0,54	-0,39	0,80	1,66	0,10	0,39
29	-0,51	-0,37	0,83	1,71	0,13	0,42
30	-0,48	-0,34	0,86	1,75	0,15	0,44
31	-0,45	-0,32	0,89	1,79	0,18	0,46
32	-0,42	-0,30	0,92	1,82	0,21	0,49

Tabel D.1: Percentiler *(continued)*

Percentil	2017	DNT	2017	DNT	2017	DNT
33	-0,39	-0,27	0,94	1,86	0,23	0,51
34	-0,37	-0,25	0,97	1,89	0,26	0,53
35	-0,34	-0,23	1,00	1,92	0,28	0,55
36	-0,31	-0,21	1,03	1,95	0,31	0,57
37	-0,29	-0,19	1,06	1,99	0,34	0,59
38	-0,26	-0,17	1,08	2,02	0,36	0,61
39	-0,23	-0,14	1,11	2,05	0,39	0,64
40	-0,21	-0,12	1,14	2,09	0,42	0,66
41	-0,18	-0,10	1,17	2,12	0,44	0,68
42	-0,16	-0,08	1,20	2,15	0,47	0,70
43	-0,13	-0,07	1,22	2,19	0,50	0,73
44	-0,11	-0,05	1,25	2,22	0,52	0,75
45	-0,08	-0,03	1,28	2,25	0,55	0,78
46	-0,06	-0,01	1,30	2,27	0,58	0,81
47	-0,04	0,01	1,33	2,30	0,60	0,84
48	-0,01	0,03	1,35	2,33	0,63	0,86
49	0,01	0,05	1,38	2,36	0,66	0,89
50	0,03	0,07	1,40	2,39	0,68	0,92
51	0,05	0,08	1,43	2,41	0,71	0,95
52	0,08	0,11	1,45	2,44	0,74	0,98
53	0,10	0,13	1,48	2,47	0,77	1,01
54	0,13	0,15	1,50	2,50	0,79	1,04
55	0,15	0,17	1,53	2,52	0,82	1,07
56	0,17	0,19	1,55	2,55	0,85	1,10
57	0,20	0,21	1,58	2,58	0,87	1,13
58	0,22	0,24	1,60	2,60	0,90	1,16
59	0,25	0,26	1,63	2,63	0,93	1,19
60	0,27	0,28	1,65	2,66	0,95	1,22
61	0,30	0,31	1,68	2,69	0,98	1,25
62	0,33	0,33	1,70	2,71	1,01	1,28
63	0,35	0,35	1,73	2,73	1,03	1,31
64	0,38	0,38	1,75	2,76	1,06	1,34
65	0,40	0,40	1,77	2,78	1,09	1,37
66	0,43	0,42	1,80	2,81	1,11	1,40
67	0,45	0,45	1,83	2,84	1,14	1,44
68	0,48	0,47	1,86	2,86	1,17	1,47
69	0,51	0,49	1,88	2,89	1,20	1,50
70	0,53	0,52	1,91	2,92	1,22	1,54
71	0,56	0,54	1,94	2,95	1,25	1,58
72	0,59	0,57	1,97	2,98	1,28	1,61
73	0,62	0,59	2,00	3,01	1,31	1,65
74	0,64	0,62	2,02	3,03	1,34	1,68
75	0,67	0,65	2,05	3,06	1,37	1,73
76	0,71	0,68	2,08	3,09	1,41	1,77
77	0,74	0,71	2,11	3,12	1,44	1,80
78	0,77	0,75	2,14	3,14	1,47	1,84

Tabel D.1: Percentiler *(continued)*

Percentil	2017	DNT	2017	DNT	2017	DNT
79	0,81	0,79	2,17	3,18	1,51	1,88
80	0,84	0,83	2,21	3,22	1,54	1,92
81	0,87	0,88	2,25	3,26	1,58	1,97
82	0,91	0,93	2,29	3,29	1,62	2,02
83	0,96	0,98	2,32	3,34	1,66	2,06
84	1,00	1,05	2,36	3,38	1,70	2,11
85	1,05	1,12	2,40	3,42	1,75	2,16
86	1,10	1,21	2,44	3,47	1,79	2,21
87	1,16	1,29	2,49	3,53	1,84	2,27
88	1,23	1,40	2,54	3,58	1,89	2,33
89	1,30	1,55	2,59	3,65	1,95	2,40
90	1,38	1,70	2,65	3,69	2,01	2,46
91	1,49	1,86	2,72	3,78	2,07	2,54
92	1,61	2,07	2,80	3,87	2,14	2,62
93	1,72	2,28	2,87	3,93	2,21	2,70
94	1,87	2,43	2,96	4,02	2,29	2,79
95	2,03	2,61	3,07	4,16	2,41	2,92
96	2,19	2,79	3,19	4,27	2,53	3,05
97	2,39	2,98	3,34	4,47	2,67	3,20
98	2,63	3,20	3,54	4,67	2,87	3,41
99	3,07	3,63	3,99	5,32	3,18	3,75
100	5,58	6,73	6,86	7,00	5,17	6,24

E

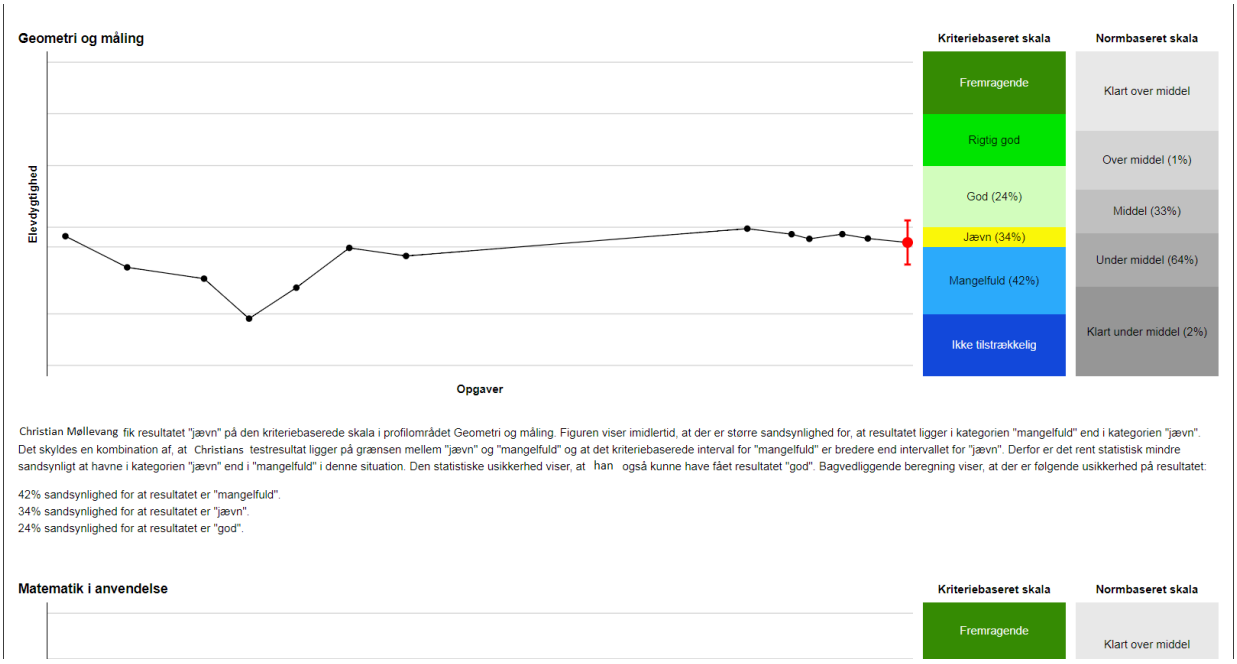
Ministeriets visning af kriteriebaserede scores

I dette bilag vises nogle eksempler på ministeriets brug af kriteriebaserede scores. Figurerne stammer fra Styrelsen for It og Læring (2016a) og Styrelsen for It og Læring (2017).

Afkodning

Betegnelser for niveauer	Beskrivelse af fagligt niveau	Inspiration til fagligt løft
Fremragende præstation	Eleven forventes at kunne afkode ord, der optræder i tekster, der anvendes på klassetrinnet	Eleven kan styrke sin sikkerhed i afkodning ved at arbejde med at øge såvel læsehastighed som ordforråd.
Rigtig god præstation	Eleven forventes at kunne afkode ord, der optræder i tekster, der anvendes på klassetrinnet	Eleven kan styrke sin sikkerhed i afkodning ved at arbejde med at øge såvel læsehastighed som ordforråd.
God præstation	Eleven forventes at kunne afkode ord, der optræder i tekster, der anvendes på klassetrinnet	Eleven kan styrke sin sikkerhed i afkodning ved at arbejde med at øge såvel læsehastighed som ordforråd.
Jævn præstation	Eleven forventes at kunne afkode ord, der optræder i tekster, der anvendes på klassetrinnet.	Eleven kan understøttes i at styrke sin sikkerhed i afkodning ved at arbejde med at øge såvel præcision i højtlæsning som ordforråd. Eksplicit respons på elevens afkodning kan hjælpe eleven til at udvikle strategier til kontrol af egen læsning.
Mangelfuld præstation	Eleven forventes at kunne afkode ord, der optræder i tekster, der anvendes på klassetrinnet. Eleven har dog af og til brug for, at tekster og enkeltord bliver læst højt.	Elevens afkodning bør udredes yderligere mhp. at vurdere, hvordan eleven bedst muligt understøttes i forståelsen af sammenhængen mellem bogstav og lyd, herunder også overvejelser om, hvordan anvendelse af it-teknologi i undervisningen kan understøtte elevens læring. I den daglige undervisning skal muligheden for højtlæsning af såvel tekster som enkeltord altid være tilstede i alle fag.
Ikke tilstrækkelig præstation	Eleven forventes at være usikker i at afkode ord, der optræder i tekster, der anvendes på klassetrinnet. Eleven forventes at have brug for, at tekster og enkeltord bliver læst højt.	Elevens afkodning bør udredes yderligere mhp. at vurdere, hvordan eleven bedst muligt understøttes i forståelsen af sammenhængen mellem bogstav og lyd, udvikle strategier til kontrol af egen læsning og ikke mindst, hvordan anvendelse af it-teknologi kan understøtte elevens læring både i dagligdagen og i undervisningen i alle fag.

Figur E.1 Beskrivelser af de faglige niveauer (for afkodning)



Figur E.2 Visning af testforløb og statistisk usikkerhed i forhold til de kriteriebaserede scores.

F

Georg Breddams notat om nationale test

De følgende sider er gengivet med tilladelse fra Georg Breddam og opsummerer de iagttagelser han havde gjort sig før henvendelsen til os.

De nationale tests set fra lærerperspektiv.

Starten for min interesse for DNT ligger år tilbage ved en frivillig test. Jeg studsede over et par elevresultater. Ved nærmere kig på dem, rejste sig bl.a. spørgsmålet omkring god/dårlig start og "fastholdelse" på tidligt estimeret niveau?

Jeg skrev et brev til styrelsen for Undervisning og Kvalitet og fik svaret, at alt var i orden i forhold til det adaptive princip. Særligt den afsluttende bemærkning i svaret vakte min nysgerrighed:

Kort sagt: Dine elevers testforløb ser, i lyset af det adaptive princip, fine ud. De er dog begge eksempler på elever, der enten starter rigtig dårligt eller rigtig godt.

Jeg håber, at du kan bruge mine svar.

Mange hilsner,
xxxxxx

Ifølge min opfattelse af et adaptivt testforløb burde starten være mindre væsentlig.

Spørgsmål der rejste sig:

- Er testen adaptiv i "folkelig forståelse"?
- Når testen beskrives som adaptiv/tilpasser sig den enkelte elevs niveau, er det for mig påfaldende, hvor meget **en god/dårlig start** sætter sit præg langt hen i testforløbet i de enkelte profilområder - **stiafhængighed**?
- Hvorledes/hvordan kan testen kategorisere forskellige elever med samme svarmønster vidt forskelligt ved at **stoppe forskellige steder i forløbet** af items? Nogle elever får mange spørgsmål - mens andre "godkendes" med væsentlig færre spørgsmål i samme profilområde?

Ovenstående gav anledning til et forsøg i den frivillige test. Mine elever fik hjælp til at svare i forudbestemte svarmønstre. God/dårlig start - overspring af opgaver - osv. Blot for at teste testen.

Gennemgik resultaterne, og de underbyggede min fornemmelse. Siden da er jeg flere gange stødt på den samme problematik, som ovenstående spørgsmål er udtryk for.

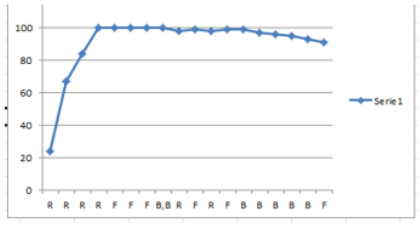
Efter flere års "frustration" måtte mine antagelser stå sin prøve. Efter artikel i Folkeskolen 7.nov 2018 omkring anbefalinger til evaluering af De nationale Tests tog jeg chancen og kontaktede Jeppe Bundsgård, som en ekspert på området. Vi fik et møde i stand og jeg fremlagde mine resultater og spørgsmål. Jeppe Bundsgård var lydhør og interesseret og gik derefter i gang med hans egne analyser med et videnskabeligt udgangspunkt.

1. "Stiafhængighed"?

Eksempel på god start og lang fastholdelse i høj sværhedsgrad:

Et eksempel på Tal og Algebra-sekvens.

Spørgsmål nr.	Sværhed i %	Rigtigt/ Forkert
1	24	R
4	67	R
7	84	R
10	100	R
13	100	F
16	100	F
19	100	F
22	100	B,B
25	98	R
27	99	F
29	98	R
31	99	F
33	99	B
35	97	B
37	96	B
39	95	B
40	93	B
41	91	F

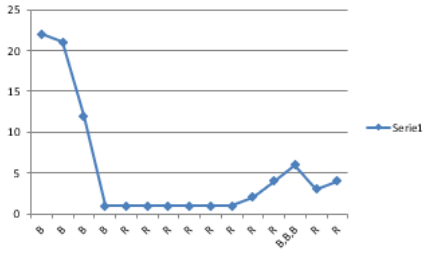


6 rigtige - 6 forkerte - 7 oversprungne
Bedømmelse: Rigtig god. (Fremragende-Rigtig god-God-Jævn-Mangelfuld-Ikke tilfredsstillende)

Eksempel på dårlig start (B er oversprunget spørgsmål = fejl) og lang fastholdelse i minimal sværhedsgrad på trods af mange rigtige:

Et eksempel på Tal og Algebra-sekvens.

Spørgsmål nr.	Sværhed i %	Rigtigt/ Forkert
1	22	B
4	21	B
7	12	B
10	1	B
13	1	R
16	1	R
19	1	R
22	1	R
25	1	R
28	1	R
31	2	R
34	4	R
37	6	B,B,B
45	3	R
48	4	R



10 rigtige - 0 forkerte - 7 oversprungne
Bedømmelse: Mangelfuld. (Fremragende-Rigtig god-God-Jævn-Mangelfuld-Ikke tilfredsstillende)

Læg mærke til midterste kolonne, hvor sværhedsgraden af de enkelte spørgsmål fremgår på en skala fra 1-100. Ikke store udsving i forhold til den - efter 3-4 startspørgsmål estimerede elevdygtighed.

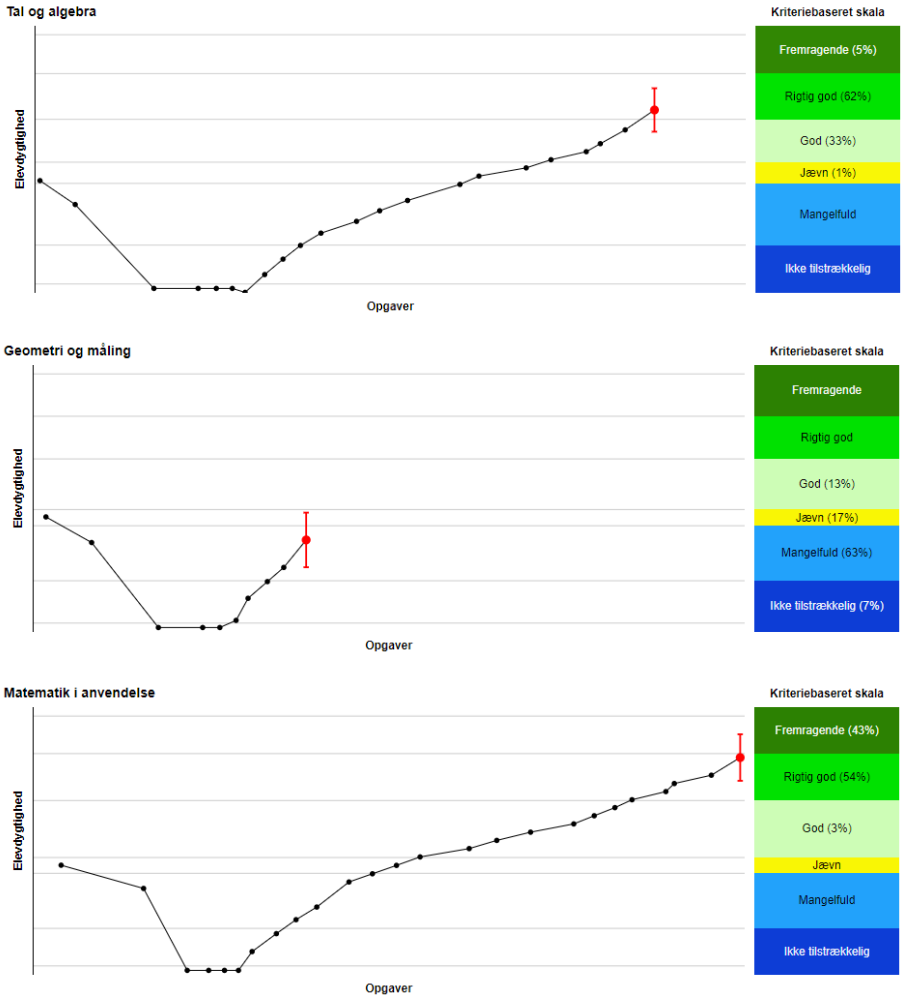
Figur F.1 Enhederne på figurerne er percentilscores på y-aksen og itemnummer i elevens forløb på x-aksen (JB og SK).

2. Tilfældige stop i testforløb:

Nedenstående elev har svaret forkert på de 4 første spørgsmål i hver af de 3 profilområder. Derefter er alle rigtige.

Hvorfor stopper testen langt før i geometridelen end i de 2 andre? (Prikker er spørgsmål) Alle 3 kurver er opadgående. Har eleven nået sit niveau efter en dårlig "nervøs" start??

Figurerne på denne side viser xx's testforløb i hvert enkelt profilområde. Den røde prik viser xx's testresultat. Der er imidlertid statistisk usikkerhed forbundet med alle test, og derfor kan det ikke udelukkes, at xx's testresultat kunne være lidt højere eller lavere. Den røde linje viser den statistiske usikkerhed i form af et interval, som xx's testresultat med stor sandsynlighed ligger indenfor.



Figur F.2 Figurerne stammer fra ministeriets visning af resultater til lærerne. Enhederne på figurerne er formentlig rasch-scores på y-aksen og itemnummer i elevens forløb på x-aksen (JB og SK).

F.0 143

88

G

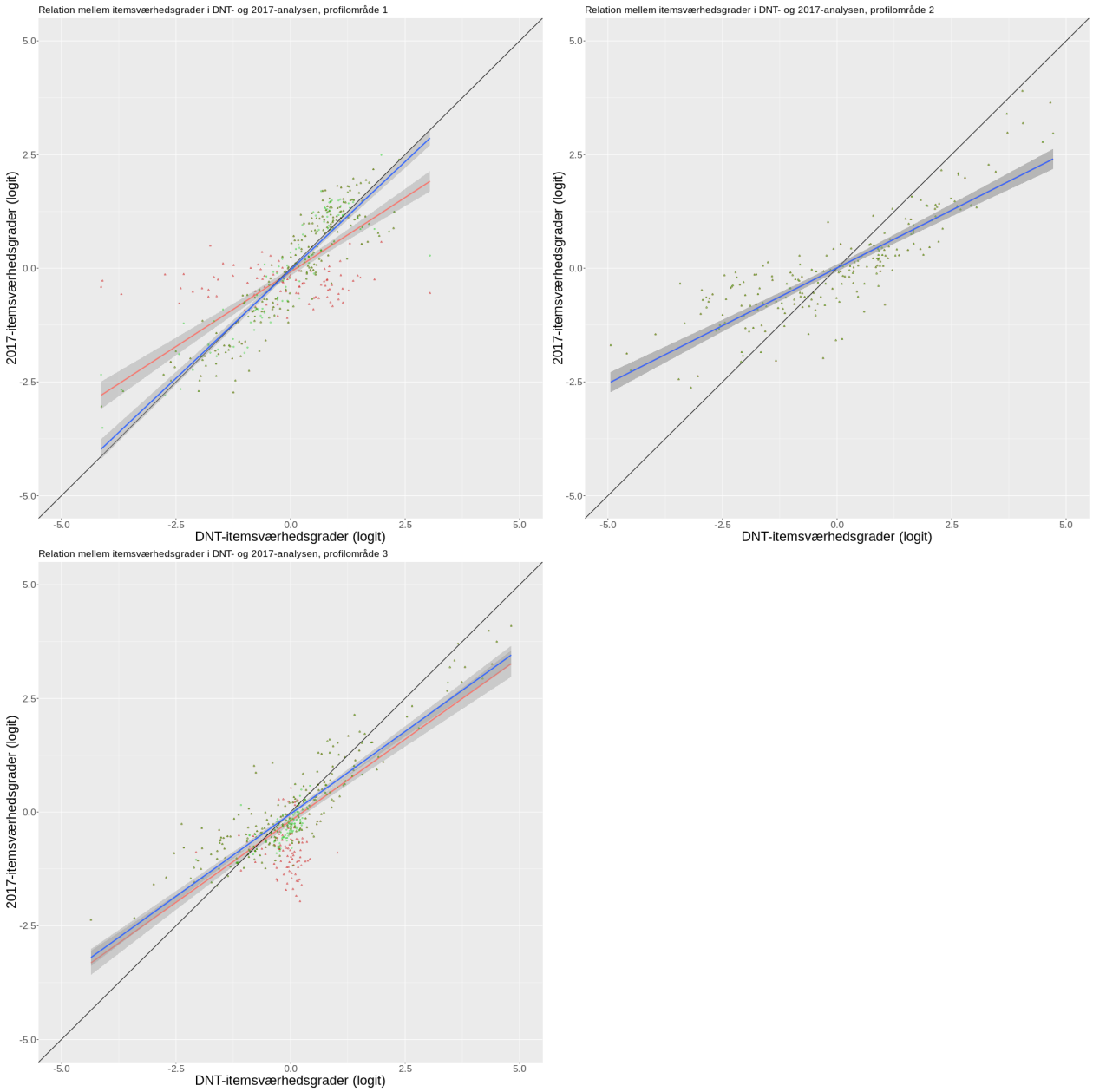
Bilag. Gennemgang af konsekvenser af fejl ved beregning af itemsværhedsgrader

Morten Rasmus Puck opdagede ved læsning af denne rapport at der var en række fejl i den tabel over itemsværhedsgrader der er gengivet i bilag C. Ved gennemgang af vores scripts til produktion af de data der indgår i tabellen, opdagede vi at der var sket en fejl som betød at item-location for polytome items (som kun findes i profilområde 3) var blevet indskrevet i profilområde 1 og derved havde overskrevet de korrekte sværhedsgrader. Det betød at 90 items havde fået forkerte sværhedsgrader i profilområde 1, og at 90 items i profilområde 3 var sat lig med den første kategoris threshold. Profilområde 2 er ikke påvirket af fejlen.

Fejlen har alene konsekvenser for de figurer der gengiver forskelle i itemsværhedsgrader mellem 2017-analysen og DNT-analysen. I alle øvrige analyser har vi ikke anvendt disse omregnede værdier.

I det følgende viser vi med samme figurer som i kapitel 4 hvad denne fejl konkret betyder for vores analyser. Sidst i bilaget gengiver vi en tabel der sammenligner alle forkerte og korrekte værdier.

Som det fremgår af figur G.1 er der ingen forskel mellem itemsværhedsgrader i profilområde 2. I profilområde 3 betød fejlen at en række items primært med sværhedsgrader omkring 0 logit var blevet givet sværhedsgrader der for de flestes vedkommende lå under 0, og for en dels vedkommende langt under 0. For profilområde 1, som jo ved fejlen havde fået tildelt forkerte sværhedsgrader, er der tale om items som ifølge DNT's analyser ligger over hele spektret, men som ved fejlen for de flestes vedkommende blev givet sværhedsgrader tættere på 0 logit.



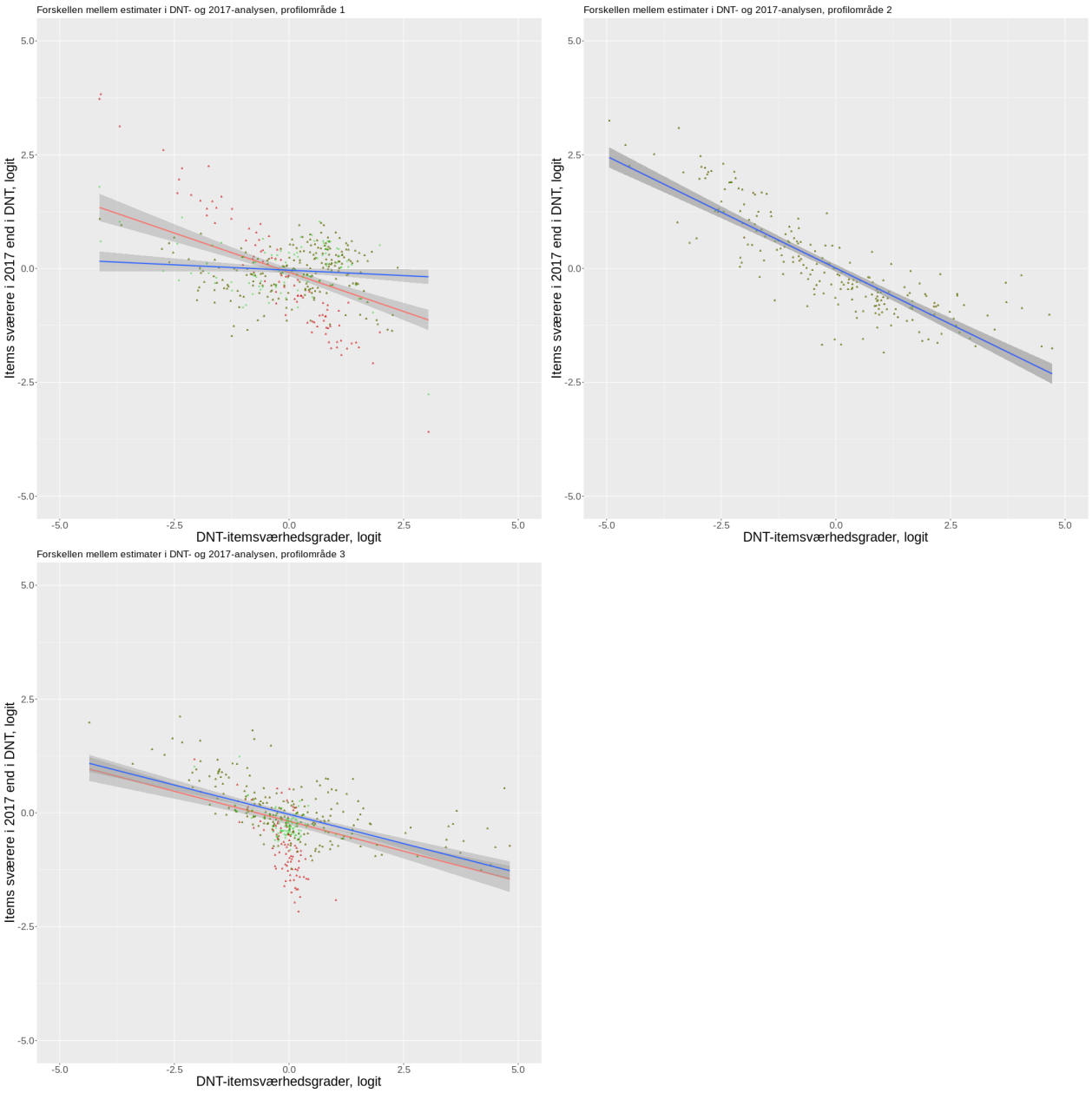
Figur G.1 Sammenligning af den korrekte og den forkerte sammenligning af 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder. Der er indtegnet en sort identitetslinje og en blå linje med konfidensinterval der viser den korrekte regression af 2017-værdier på DNT-værdier. Med rød linje er indtegnet den forkerte regression med konfidensinterval. De grønne prikker er de korrekte itemsværhedsgrader, mens de røde trekanter er de forkerte itemsværhedsgrader. De items der ikke har fået ændret deres sværhedsgrad, fremstår mørkere (fordi der både er en grøn prik og en rød trekant).

Figur G.2 viser forskellen på den korrekte og den forkerte analyses identifikation af forskelle mellem DNT's og 2017-analysens estimat af itemsværhedsgrader plottet som funktioner af DNT-sværhedsgraderne.

Som det fremgår har de forkerte itemsværhedsgrader haft en stor betydning for hvordan fordelingen af uoverensstemmelser mellem 2017-analysen og DNT-analysens itemsværhedsgrader ser ud. Hvor de forkerte sværhedsgrader gav indtryk af at det særligt var de lette items der var blevet sværere, og de svære der

var blevet lettere, giver den korrekte analyse indtryk af at der er tale om en jævnt fordelt uenighed om sværhedsgraderne.

For profilområde 3 er hældningen den samme, og både de forkerte og de korrekte forskelle på itemsværhedsgrader støtter en konklusion om at lette items i profilområde 3 er blevet sværere, mens svære items er blevet lettere. Forskydningen af den korrekte regressionslinje opad viser at der dog ikke er tale om at de er blevet helt så meget lettere over en bred kam som den forkerte analyse angav.



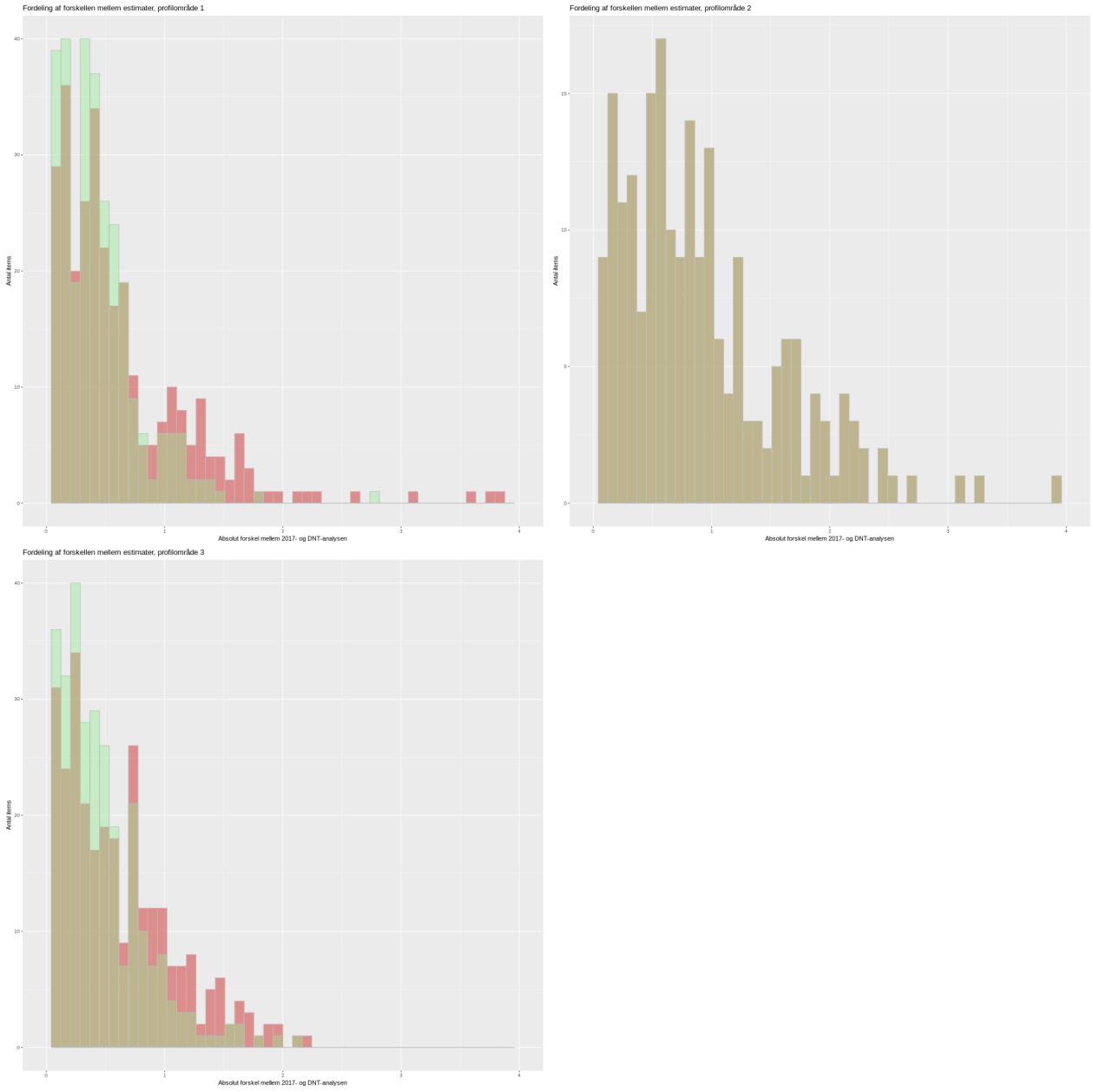
Figur G.2 Sammenligning af den korrekte og den forkerte analyses identifikation af forskelle mellem 2017-analysens og DNT's estimat af sværhedsgrader for de tre profilområder i forhold til sværhedsgraden i DNT. Der er indtegnet en blå linje med konfidensinterval der viser den korrekte regression af forskellen på 2017-værdier og DNT-værdier. Med rød linje er indtegnet den forkerte regression af forskellene med konfidensinterval. De grønne prikker er de korrekte forskelle i itemsværhedsgrader, mens de røde trekanter er de forkerte forskelle i itemsværhedsgrader. De items der ikke har fået ændret deres sværhedsgrad, fremstår mørkere (fordi der både er en grøn prik og en rød trekant).

Histogrammerne i figur G.3 viser forskellen mellem den forkerte og den korrekte analyses fordeling af den absolutte forskel på de to estimater af sværhedsgraden for de tre profilområder.

Som det fremgår er der generelt tale om at der er færre meget store forskelle i den korrekte analyse i både profilområde 1 og 3. Men der er dog stadig tale om at 96 items udviser en forskel der er større end 0,5 logit

i profilområde 1, mens det drejer sig om 99 items der udviser en forskel større end 0,5 logit i profilområde 3. For profilområde 2 er det fortsat 149 items der udviser forskelle over 0,5 logit.

Den korrekte gennemsnitlige absolutte forskel er henholdsvis 0,42, 0,91 og 0,44, mens den forkerte gennemsnitlige absolutte forskel var henholdsvis 0,61, 0,91 og 0,59



Figur G.3 Forskellen på fordelingen af forskelle mellem 2017-analysens og DNT's estimater af sværhedsgrader for de tre profilområder. De røde søjler viser de forkerte beregninger, mens de grønne viser de korrekte. De brune angiver der hvor der er overlap mellem de to analyser.

G.1 Vurdering af konsekvenser af de forkerte beregninger

Som det fremgår af figurerne og tabellen (tabel G.1) i næste afsnit, har de forkerte itemsværhedsgrader generelt betydet at forskellene så større ud end de var. Men som det også fremgår, så ændrer fejlen ikke på at der er tale om meget store forskelle mellem DNT's og den korrekte 2017-analyses estimat af itemsværhedsgrader. Således gælder det for alle tre profilområder at mere end 90 items udviser forskelle der er større end 0,5 logit. Vi ser således ingen grund til at ændre de konklusioner som vi har foretaget i rapporten.

G.2 Tabel over forkerte og korrekte værdier

Tabel G.1: Oversigt over forkerte og korrekte itemsværhedsgrader.

Item	Korrekt værdi	Forkert værdi	Forskel i værdi
0108010003012389 25-1	1.35	-0.02	1.36
0108010003012389 94-1	-1.04	-0.03	-1.02
0108010003012389 95-1	-0.26	-0.51	0.24
0108010003012389 96-1	0.28	-0.06	0.34
0108010003012389 99-1	0.94	-0.07	1.01
0108010003012390 00-1	1.45	-0.28	1.73
0108010003012391 95-1	1.32	-0.39	1.71
0108010003012392 35-1	-0.11	-0.08	-0.03
0108010410080	-2.66	-0.44	-2.21
0108010410088	1.30	-0.25	1.55
0108010410093	-1.36	-0.58	-0.77
0108010410094	0.86	-0.25	1.11
0108010410098	-1.74	-0.32	-1.42
0108010410110-1	-0.89	-0.24	-0.65
0108010410145-1	1.22	-0.26	1.48
0108010410155-1	0.35	-0.16	0.50
0108010410186-1	-3.51	-0.28	-3.23
0108010410187-1	-1.09	0.36	-1.44
0108010410230005	0.13	-0.33	0.46
0108010410230008	-0.29	-0.40	0.12
0108010410230009	0.08	-0.41	0.50
0108010410230012	1.02	-0.29	1.31
0108010410230013	0.42	-0.26	0.68
0108010410230018	-0.72	-0.25	-0.47
0108010410230019	0.74	-0.33	1.07
0108010410230020	-0.18	-0.80	0.63
0108010410230021	-0.59	-0.22	-0.37
0108010410230022	0.11	-0.55	0.66
0108010410230023	0.95	-0.34	1.29
0108010410230024	1.27	-0.46	1.73

Tabel G.1: Oversigt over forkerte og korrekte itemsværhedsgrader. *(continued)*

Item	Korrekt værdi	Forkert værdi	Forskel i værdi
0108010410230025	-1.56	-0.17	-1.39
0108010410230028	0.21	-0.30	0.50
0108010410230029	-0.98	-0.17	-0.81
0108010410230030	-0.28	-0.31	0.04
0108010410230031	1.16	-0.63	1.79
0108010410230032	-0.90	0.08	-0.98
0108010410230034	1.70	-0.37	2.06
0108010410230035	1.16	-0.45	1.61
0108010410230037	1.49	-0.37	1.86
0108010410230038	0.52	-0.56	1.08
0108010410230039	-1.19	0.27	-1.46
0108010410230041	-0.08	-0.14	0.06
0108010410230042	-1.69	-0.63	-1.07
0108010410230045	-0.91	0.10	-1.01
0108010410230046	-0.22	-0.20	-0.03
0108010410230047	1.49	-0.69	2.18
0108010410311	-2.34	-0.41	-1.93
0108010410315	0.28	-0.55	0.83
0108010410316	-1.21	-0.13	-1.08
0108010410320	0.04	0.16	-0.12
0108010410325	-1.19	0.00	-1.20
0108010410327	-1.16	-0.20	-0.97
0108010410328	1.45	0.04	1.41
0108010410333	-0.87	-0.85	-0.02
0108010410335	-1.67	-0.48	-1.19
0108010410337	-1.94	-0.44	-1.50
0108010410339	-1.79	-0.25	-1.54
0108010410340	-2.79	-0.14	-2.65
0108010410343	-1.89	-0.78	-1.11
0108010410344	-2.67	-0.57	-2.10
0108010410350	1.10	-0.16	1.26
0108010410351	-0.69	-0.80	0.11
0108010410357	-0.66	-0.41	-0.26
0108010410358	-1.86	-0.93	-0.93
0108010410372	0.86	-0.84	1.69
0108010410373	0.84	-0.66	1.50
0108010410376	-1.86	0.49	-2.35
0108010410385	-1.93	-0.62	-1.32
0108010410388	1.17	-0.57	1.74
0108010410392	-0.12	-0.65	0.54
0108010410393	0.11	-0.79	0.90
0108010410400-1	-0.64	-1.05	0.41
0108010410405-1	1.17	-0.76	1.93
0108010410407-1	0.86	-0.21	1.07
0108010410414-1	-0.72	0.01	-0.73

Tabel G.1: Oversigt over forkerte og korrekte itemsværhedsgrader. *(continued)*

Item	Korrekt værdi	Forkert værdi	Forskel i værdi
0108010415102	-2.24	-0.52	-1.72
0108010415109	1.43	-0.49	1.92
0108010415111	0.45	-0.06	0.51
0108010415113	1.33	0.54	0.78
0108010415180	2.49	0.58	1.91
0108010420003	-0.41	-0.51	0.10
0108010420041	-1.73	0.07	-1.80
0108010420042	-0.93	-0.51	-0.42
0108010420088	1.13	-0.64	1.77
0108010420129	1.51	0.16	1.36
0108010420138	1.45	-0.75	2.20
0108010420141	0.06	-1.10	1.16
0108010420152	0.88	-0.49	1.36
0108010420153	1.37	-0.15	1.52
0108010420157	1.23	-0.91	2.14
010803000301238567-1	0.48	0.48	0.00
010803000301238628-1	1.52	1.52	0.00
010803000301238640-1	-0.03	-1.49	1.46
010803000301238646-1	-0.51	-0.76	0.25
010803000301238708-1	1.28	1.28	0.00
010803000301238728-1	-0.07	-0.61	0.54
010803000301238729-1	-0.28	0.53	-0.81
010803000301238841-1	0.89	0.89	0.00
010803000301238844-1	-0.08	-1.52	1.44
010803000301238868-1	1.45	1.45	0.00
010803000301238871-1	-0.25	-1.10	0.85
010803000301238874-1	-0.58	-0.61	0.03
010803000301238989-1	0.22	0.22	0.00
010803000301238990-1	-0.32	0.28	-0.60
010803000301238992-1	-0.24	0.23	-0.47
010803000301239051-1	-0.26	-0.88	0.61
010803000301239061-1	-0.16	-0.48	0.33
010803000301239062-1	-0.28	0.28	-0.55
010803000301239063-1	0.36	-1.96	2.32
010803000301241375-1	-0.33	0.27	-0.60
010803000301241378-1	-0.40	-1.07	0.67
010803000301241381-1	-0.41	-1.69	1.28
010803000301241383-1	-0.29	-1.61	1.32
010803000301241391-1	-0.01	-0.01	0.00
010803000301241409-1	-0.25	-0.58	0.33
010803000301241410-1	-0.33	-0.36	0.02
010803000301241412-1	-0.80	-0.73	-0.07
010803000301241413-1	-0.22	-0.84	0.62
010803000301241414-1	-0.55	-0.28	-0.27
010803000301241415-1	-0.34	-1.37	1.03

Tabel G.1: Oversigt over forkerte og korrekte itemsværhedsgrader. *(continued)*

Item	Korrekt værdi	Forkert værdi	Forskel i værdi
010803000301241593-1	-0.46	-0.97	0.51
010803000301241597-1	-0.17	-1.00	0.83
010803000301241806-1	-0.30	-1.46	1.17
010803000301241807-1	-0.17	-1.52	1.34
010803000301241808-1	-0.31	-0.49	0.17
010803000301241811-1	-0.63	-0.48	-0.16
010803000301241812-1	0.08	-1.84	1.92
010803000301241917-1	-0.37	-0.59	0.23
010803000301241921-1	-0.45	-0.72	0.27
010803000301241924-1	-0.37	-0.60	0.23
010803000301241925-1	-0.56	-0.51	-0.05
010803000301241926-1	0.27	-1.09	1.36
010803000301242169-1	-0.14	-0.63	0.49
010803000301242171-1	-0.63	-1.21	0.59
010803000301242180-1	0.10	-0.66	0.76
010803000301242204-1	-0.20	-0.84	0.65
010803000301242207-1	-0.69	-0.43	-0.26
010803000301242228-1	-0.41	-0.48	0.07
010803000301242234-1	-0.55	-1.37	0.83
010803000301242250-1	-0.13	-0.67	0.54
010803000301242254-1	0.16	-0.68	0.84
010803000301242264-1	0.00	-1.10	1.10
010803000301242269-1	-0.20	-0.80	0.60
010803000301242272-1	0.04	-0.34	0.38
010803000301242274-1	-0.85	-0.82	-0.04
010803000301242310-1	-0.48	-0.12	-0.35
010803000301242313-1	-0.44	-1.08	0.64
010803000301242315-1	-0.25	-0.91	0.66
010803000301242596-1	-0.14	-1.30	1.16
010803000301242597-1	-0.78	-0.49	-0.29
010803000301242601-1	-0.57	-1.71	1.14
010803000301242603-1	-0.16	-1.17	1.01
010803000301242744-1	-0.80	-0.75	-0.04
010803000301243169-1	-0.41	-0.27	-0.14
01080303060910323	-0.44	-0.44	0.00
01080303060910325	-0.84	-0.82	-0.02
01080303060910327	-0.66	-0.93	0.27
0108030310009-2	0.68	0.68	0.00
0108030310011-1	-0.62	-1.51	0.90
0108030310015-2	-0.57	-1.57	1.01
0108030310016-2	-0.65	-1.14	0.49
0108030310372	0.82	0.82	0.00
0108030310609-1	1.02	1.02	0.00
0108030310612-1	-0.48	-0.48	0.00
0108030311028	-0.99	-0.99	0.00

Tabel G.1: Oversigt over forkerte og korrekte itemsværhedsgrader. *(continued)*

Item	Korrekt værdi	Forkert værdi	Forskel i værdi
0108030311030	0.01	-1.21	1.22
0108030320015	0.29	0.29	0.00
0108030320016	-0.49	-1.10	0.61
0108030320019-2	-0.06	-0.85	0.79
0108030320053-1	0.84	0.84	0.00
0108030320063-1	-0.26	-0.26	0.00
01080306061330063-2	-0.39	-0.39	0.00
01080306061330065-1	0.07	-1.07	1.14
01080306061340006-1	0.28	0.28	0.00
0108030610700-2	-0.61	-0.61	0.00
0108030610705-2	-0.43	-0.43	0.00
0108030612006-1	-1.46	-1.46	0.00
0108030612014-1	0.59	0.59	0.00
0108030612014-2	-0.49	-0.81	0.32
0108030612019-2	-0.05	-0.05	0.00

Litteratur

Adams, Raymond J., og Mark Wilson. 1996. "Formulating the Rasch Model as a Mixed Coefficients Multinomial Logit Model". I *Objective Measurement: Theory into Practice Volume 3*, 143–66. Norwood, NJ: Ablex Publishing.

Adams, Raymond J., og Margaret L. Wu. 2007. "The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model". I *Multivariate and Mixture Distribution Rasch Models*, redigeret af Matthias von Davier og Claus H. Carstensen, 57–76. New York: Springer Science Business Media.

Adams, Raymond J., Mark Wilson, og Wen-chung Wang. 1997. "The Multidimensional Random Coefficients Multinomial Logit Model". *Applied Psychological Measurement* 21 (1): 1–23.

"Aftale Mellem Regeringen (Socialdemokraterne, Radikale Venstre Og Socialistisk Folkeparti), Venstre Og Dansk Folkeparti Om et Fagligt Løft Af Folkeskolen". 2013.

Boch, R.D., og M. Aitken. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm". *Psykometrika*, nr. 46: 443–59.

Bundsgaard, Jeppe. 2016. "Rapport Om Kompetencetest i Hitte På-Projektet". København: DPU, Aarhus Universitet.

———. 2018a. "Det 21. Århundredes Kompetencer". I *Skoleudvikling Med It: Forskning i Tre Demonstrationsskoleforsøg I*, redigeret af Jeppe Bundsgaard, Marianne Georgsen, Stefan Graf, og Thomas Illum Hansen, 143–65. Aarhus: Aarhus Universitetsforlag.

———. 2018b. "Test Og Måling i Fagene". I *Udvikling i Didaktik Didaktik i Udvikling*, redigeret af Torben Spanget Christensen, Nikolaj Elf, Peter Hobel, og Ane Qvortrup. Aarhus: Klim.

———. 2018c. "Pædagogisk brug af test". *Sakprosa* 10 (2). doi:10.5617/sakprosa.6007.

Bundsgaard, Jeppe, og Morten Rasmus Puck. 2016. *Nationale Test - Danske Lærere Og Skolelederes Brug, Holdninger Og Viden*. DPU, Aarhus Universitet.

Care, Esther, Claire Scoular, og Patrick Griffin. 2016. "Assessment of Collaborative Problem Solving in Education Environments". *Applied Measurement in Education* 29 (4): 250–64. doi:10.1080/08957347.2016.1209204.

Feinman, Joshua. 2008. "High Stakes, but Low Validity? A Case Study of Standardized Tests and Admissions

into New York City Specialized High Schools." Boulder; Tempe: Education and the Public Interest Center & Education Policy Research Unit.

Geisinger, Kurt F. 2016. "21st Century Skills: What Are They and How Do We Assess Them?" *Applied Measurement in Education* 29 (4): 245–49. doi:10.1080/08957347.2016.1209207.

Greenstein, Laura M. 2012. *Assessing 21st Century Skills: A Guide to Evaluating Mastery and Authentic Learning*. Corwin Press.

Griffin, Patrick, og Esther Care, red. 2015. *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-9395-7.

Hale, Charles Dennis, og Douglas Astolfi. 2014. *Measuring Learning and Performance: A Primer*. 3rd edition. St. Leo, Florida: Saint Leo University.

Harbo, Ulf. 2015. "Norddjurs kommune udfordrer de Nationale test - Folkeskolen.dk". https://www.folkeskolen.dk/573670/norddjurs-kommune-udfordrer-de-nationale-test.

Hoover, Wesley A., og Philip B. Gough. 1990. "The Simple View of Reading". *Reading and writing* 2 (2): 127–60.

Kousholt, Kristine. 2015. "Børns Gætterier Ved Nationale Test". *Cepra-striben*, Nationale Tests„ nr. 18: 46–57.

Kreiner, Svend. 2007. "Den Adaptive Procedure".

Kreiner, Svend, og Karl Bang Christensen. 2013. "Person Parameter Estimation and Measurement in Rasch Models". I *Rasch Models in Health*, redigeret af Karl Bang Christensen, Svend Kreiner, og M Mesbah, 63–78. London: ISTE & John Wiley & Sons.

"Lov Om Ændring Af Lov Om Folkeskolen". 2006.

Norling, Marina. 2016. "De nationale test 2016 - hvor galt står det til? (1) - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/586828/de-nationale-test-2016—hvor-galt-staar-det-til-1.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in Mathematical Psychology, Vol. 1. Copenhagen: Danmarks Pædagogiske Institut.

Ravn, Karen. 2014. "Ups de nationale test måler ikke så præcist som lovet - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/539694/ups–de-nationale-test-maaler-ikke-saa-praecist-som-lovet.

———. 2015a. "Duer ikke: En femtedel af opgaverne i de nationale test kasseret - Folkeskolen.dk". *folkeskolen.dk*. https://www.folkeskolen.dk/572751/duer-ikke-en-femtedel-af-opgaverne-i-de-nationale-test-kasseret.

———. 2015b. "Eksperter dumper de nationale test - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/572813/eksperter-dumper-de-nationale-test.

———. 2015c. "Skoleleder: Testresultater svinger, som vinden blæser - Folkeskolen.dk". *Folkeskolen.dk*. https://www.folkeskolen.dk/572808/skoleleder-testresultater-svinger-som-vinden-blaeser.

Riise, Andreas Brøns. 2014. "Nationale test: Klasse havde kæmpe udsving på tre dage - Folkeskolen.dk". https://www.folkeskolen.dk/540229/nationale-test-klasse-havde-kaempe-udsving-paa-tre-dage.

Robitzsch, Alexander, Thomas Kiefer, og Margaret Wu. 2019. *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM.

Rosenbaum, Paul R. 1989. "Criterion-Related Construct Validity". *Psychometrika* 54 (4): 625–33. doi:10.1007/BF02296400.

Siddiq, Fazilat, Perman Gochyyev, og Mark Wilson. 2017. "Learning in Digital Networks Literacy:

A Novel Assessment of Students' 21st Century Skills". *Computers & Education* 109 (juni): 11–37. doi:10.1016/j.compedu.2017.01.014.

Smarter Balanced Assessment Consortium. 2018. "Smarter Balanced Assessment Consortium: 2016-17 Technical Report".

Styrelsen for It og Læring. 2015. "Den adaptive algoritme i De Nationale Test". København: Undervisningsministeriet, Styrelsen for It og Læring.

———. 2016a. *Test- Og Prøvesystemet - De Nationale Test. Brugervejledning for Skoler*. København: Ministeriet for Børn, Unge og Undervisning.

———. 2016b. "Undersøgelse af de nationale tests reliabilitet". København: Undervisningsministeriet.

———. 2016c. "Nationale Tests Måleegenskaber". København: Undervisningsministeriet.

———. 2017. "Vejledning Til Nye Resultatvisninger i de Nationale Test Til Lærere i Alle Fag". København: Styrelsen for It og Læring.

Styrelsen for Undervisning og Kvalitet. 2016. "Undersøgelse Af Udsving i Testresultater På Ørum Skole i Norddjurs Kommune". Styrelsen for Undervisning og Kvalitet.

Taylor, Wilson L. 1953. "Cloze Procedure: A New Tool for Measuring Readability". *Journalism & Mass Communication Quarterly* 30 (4): 415–33.

Undervisningsminister Bertel Haarder (V). 2006. "L 101 Forslag til lov om ændring af lov om folkeskolen. (Styrket evaluering og anvendelse af nationale test som pædagogisk redskab samt obligatoriske prøver m.v.)."

Wandall, Jakob. 2010. "Test, prøver og evaluering i grundskolen". Powerpoints. København: IND/KU.

Wandall, Jakob, Christine Nørrelund, og Mette Dalgaard Nielsen. 2018. "Elevernes Syn På de Nationale Test". Nordic Metrics.

Wells, Craig S, og James A Wollack. 2003. "An Instructor's Guide to Understanding Test Reliability". Testing & Evaluation Services. University of Wisconsin.

Wilson, Mark. 2003. "On Choosing a Model for Measuring". *Methods of Psychological Research-Online* 8 (3): 1–22.

———. 2005. *Constructing Measures : An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Wright, Benjamin D., og Mark H. Stone. 1979. *Best Test Design*. Chicago: MESA Press.

Wu, Margaret, Raymond J. Adams, Mark Wilson, og S. Haldane. 2007. *ConQuest: Generalised Item Response Modelling Software (Version 2.0)*. Camberwell, Australia: ACER Press.

Zwinderman, A. H. 1995. "Pairwise Parameter Estimation in Rasch Models". *Applied Psychological Measurement*, nr. 19: 369–75.