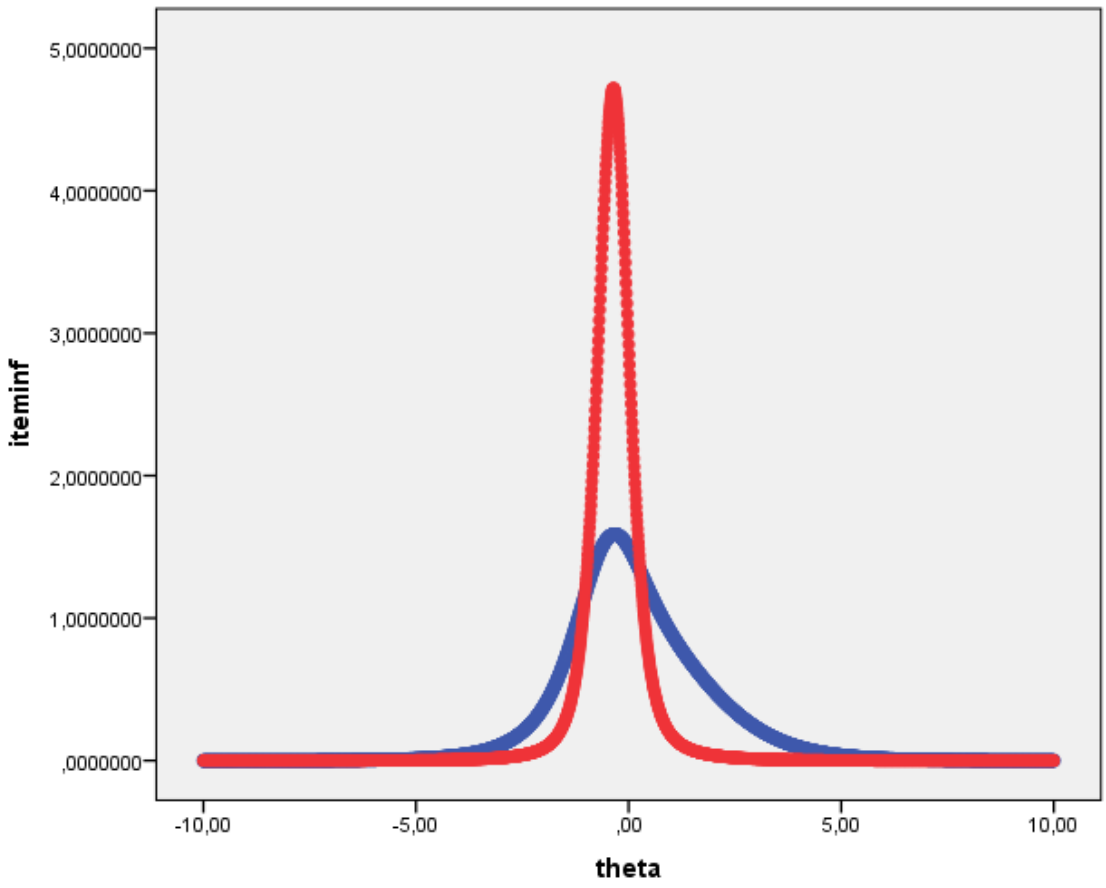
# Om opgavetyper og usikkerhed i de nationale test




Af Svend Kreiner, juni 2017

## Indholdsfortegnelse

## 1. Indledning

Styrelsen for Undervisning og Kvalitet (STUK) og Styrelsen for It og Læring (STIL) bidrager hvert år med viden til opgavekommissionerne for at sikre testopgaver til de nationale test (herefter DNT) af høj kvalitet. Visse opgavetyper er af faglige årsager mere egnet end andre i fagene, og dette vægtes højt. STUK og STIL har imidlertid bedt konsulentfirmaet Svend Kreiner om at undersøge, hvorvidt der ud fra statistiske betragtninger er nogle opgavetyper, som er mere velegnede end andre. I denne rapport foreligger konsulentfirmaets teoretiske bud på dette.

DNT skal gennemføres på 45 min. Formålet er, at bestemme elevens dygtighed i hvert af testens tre profilområder med størst mulig statistisk sikkerhed. Den statistiske usikkerhed på elevens testresultat afhænger af flere forhold. Først og fremmest skal elevens testadfærd være sådan, at svarene på opgaverne svarer til det, som Rasch-modellen forventer, idet testresultatet vil være systematisk skævt, hvis dette ikke er tilfældet. Hvis dette ikke er et problem, afhænger usikkerheden af det antal opgaver, som eleven besvarer og af den information (den såkaldte item-information), som opgaven kan bidrage med i forhold til elevens faglige niveau. Hvis item-informationerne for de opgaver, som eleven har besvaret, er høj, vil usikkerheden være begrænset.

Item-informationen afhænger af to ting. For det første af om opgavens sværhedsgrad passer til elevens niveau og for det andet af antallet af svarkategorier og af de item-parametre, der omtales som opgavernes tærskel-værdier, når der er tale om polytome items. Forholdet mellem tærskelværdierne på den ene side og den information, som opgaven kan levere, er kompliceret, og der foreligger – så vidt ministeriet eller forfatter af denne rapport ved – ingen teoretiske eller praktiske undersøgelser heraf. Det primære formål med dette projekt er derfor at undersøge sammenhængen mellem tærskel-værdier og item-information på den ene side og forskellige former for opgavetyper på den anden side for at afklare, om der er visse opgavetyper, som indebærer en højere grad af information end andre typer, og som derfor i højere grad bør anvendes i de nationale test.

Den adaptive algoritme i de nationale test vælger opgaver blandt de opgaver, hvor item-informationen er maksimeret i nærheden af elevens niveau. Det uklare forhold mellem opgavernes tærskelværdier og den maksimale information, som opgaverne kan levere, betyder, at denne måde at udvælge opgaver på ikke nødvendigvis er den optimale. Det er derfor et sekundært formål at undersøge, om der ud fra statistiske overvejelser er andre og bedre måder at vælge opgaver på, end den der i øjeblikket benyttes.

## 2. Baggrunden

DNT benytter flere forskellige opgave-formater defineret ved en kombination af opgavetype og antallet af spørgsmål, der stilles i forbindelse med opgaven.

I det efterfølgende omtales de enkelte spørgsmål, som stilles i forbindelse med en opgave som items. En DNT-opgave kan således indeholde ét item eller flere forskellige items.

I forbindelse med opgørelsen af resultaterne beregnes en opgave-score lig med antallet af items, som blev korrekt besvaret. Opgaver, som kun består af et enkelt item, omtales som *dikotome* opgaver, fordi den samlede score kun omfatter to værdier, 0 eller 1. Opgaver med flere items omtales som *polytome* opgaver, fordi den samlede opgave-score omfatter mere end to forskellige værdier.

Opgaverne i DNT antages at opføre sig som items fra såkaldte Rasch-modeller, hvor svaret på opgaverne kun antages at afhænge af en elev-parameter, der er et udtryk for elevens dygtighed, og af en eller flere opgaveparametre, der karakteriserer opgaven. Værdierne af opgaveparametrene er beregnet i forbindelse med afprøvningen af tilpasningen mellem opgaver og Rasch-modellen, hvor opgaver, der ikke passede til modellen, blev elimineret. I forbindelse med senere anvendelser af DNT, antages opgave-parametrene derfor at være kendte. Det eneste, der står tilbage, når elevernes svar på opgaverne er registreret, er derfor at estimere værdien af elev-parameteren. Det er dette estimat, der opfattes som målingen af elevens dygtighed. Da der er tale om et statistisk estimat af dygtigheden, er der – som for alle andre statistiske estimater – tale om målinger med en usikkerhed, der kan beskrives vha. estimatets standardfejl. Det er denne standardfejl som i forbindelse med pædagogiske test omtales som målingens "standard error of measurement", SEM.

## 3. Udfordringer

Standardfejl i forbindelse med statistiske estimater afhænger af, hvor stort datamateriale man har til rådighed. I almindelige statistiske undersøgelse er der som regel tale om store datamaterialer og derfor en forholdsvis beskeden standardfejl. På dette punkt afviger estimater af elev-parameteren fra almindelige statistiske estimater, fordi det datamateriale, der er til rådighed for estimationen – dvs. svarerne på de opgaver, som eleven har besvaret – er beskedent. Sammenlignet med den form for standardfejl, som man almindeligvis har at gøre med i statistiske undersøgelser, vil SEM for målinger af elevernes dygtighed være væsentlig større. Målinger ved hjælp af pædagogiske test er *altid* – uanset hvilken pædagogisk test der anvendes – målinger med usikkerhed.

Udover antallet af opgaver afhænger usikkerheden også af, hvilke opgaver eleven har besvaret. Hvis opgaverne enten er for lette eller for vanskelige for en elev, vil målingen af denne elevs dygtighed være behæftet med en meget stor usikkerhed. Det er dette problem, som man forsøger at løse ved at anvende adaptive test, hvor opgaverne vælges således, at den statistiske usikkerhed af målingen af elevdygtigheden i princippet kun kommer til at afhænge af, hvor mange opgaver eleven har besvaret, fordi eleven i mindre grad bliver bedt om at besvare opgaver, som er for lette eller for svære.

Debatten omkring DNT har blandt andet handlet om usikkerheden. Det har tilsyneladende været en overraskelse for mange, at pædagogiske testresultater er behæftet med usikkerhed, og det er blevet påstået, at usikkerheden i DNT er større end i andre pædagogiske test. Dette er rent faktisk ikke tilfældet, hvis der tales om test med det samme antal opgaver som i DNT.

Selvom det ikke er hensigten at komme nærmere ind på detaljerne i denne debat, skal det understreges, at diskussionen ofte er baseret på utilstrækkelig viden om, hvad der faktisk menes med usikkerhed, hvad de SEM størrelser, som DNT beregner, er udtryk for, og hvad SEM kan forventes at være i almindelige ikke-adaptive test. Der er flere årsager til dette. Den vigtigste er naturligvis, at vurdering af usikkerhed i pædagogiske test er et relativt kompliceret problem, men det hører også med, at den psykometriske terminologi er uhensigtsmæssig og kan være direkte vildledende. Derudover at der ikke findes andre almindeligt brugte danske pædagogiske test, hvor usikkerheden beregnes og rapporteres på den måde, som DNT gør det[1].

Uanset at diskussionen af usikkerheden i DNT i et vist omfang er behæftet med utilstrækkelig viden om dette forhold, er usikkerheden uomgængelig og en udfordring for brugerne af testene, og det er nødvendigt at gøre så meget som muligt for at reducere usikkerheden til det mindst mulige. I den forbindelse er det værd at bemærke, at der er en række specielle forhold i forbindelse med den måde opgaverne i DNT er konstrueret på, der kan påvirke usikkerheden, og at der derfor er grund til at se nærmere på, om usikkerheden i DNTs målinger af elevdygtigheden kan reduceres ved at forbedre opgaverne og/eller at ændre den måde, som DNTs adaptive rutiner udvælger opgaver på i forbindelse med testforløbet. Det er disse spørgsmål, som denne rapport vil forsøge at kaste lys over.

En af de faktorer, som komplicerer tingene, er, at DNT bruger både dikotome opgaver, hvor hver opgave kun indeholder ét item, der enten kan besvares korrekt eller forkert og polytome opgaver af den type, som teorien for adaptive test omtales som testlets[2] med to eller flere items, hvor svaret på opgaven defineres som antallet af items, der blev korrekt besvaret.

Almindelige pædagogiske test anvender kun dikotome opgaver, hvor spørgsmål om sværhedsgraden og hvor meget et svar på opgaven kan bidrage til at forbedre sikkerheden af den samlede test, forenkles i en sådan grad, at man kan bevise, at valget af opgavetyper, ikke kan have nogen konsekvenser for usikkerheden af målinger i forbindelse med adaptive test, hvis opgaverne passer til den såkaldte Rasch-model. Under forudsætning af, at der er tale om dikotome Rasch opgaver, vil adaptive test altid give mere sikre resultater end ikke-adaptive test med det samme antal opgaver, hvis de adaptive algoritmer fungerer efter hensigten, og hvis der er tilstrækkelig mange opgaver på alle niveauer af sværhedsgraden.

[1] Dette udsagn skal naturligvis forstås på den måde, at forfatteren til denne rapport ikke kender til andre danske test, hvor SEM rutinemæssigt rapporteres. En kontakt til medarbejdere på Hogrefe i forbindelse med udarbejdelsen af denne rapport bekræftede også, at de heller ikke kendte til danske test, hvor usikkerheden vurderes ved hjælp af beregninger af SEM.

[2] Testlet opgaver er beskrevet i Wainer (2000, s. 245-254). Testlet begrebet blev introduceret af Wainer og Kiely (1987, s. 190), der beskrev det på følgende måde:*" (…) a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow"*.

Så enkelt er det ikke, når det drejer sig om test med polytome opgaver. I forbindelse med sådanne test, er der grund til at stille spørgsmål, der både drejer sig om det fornuftige i at anvende polytome opgaver, og om der er visse opgavetyper, som giver mere sikre målinger end andre.

Sådanne spørgsmål er f.eks. følgende.

1) Giver DNTs polytome opgaver mere usikre målinger end et tilsvarende antal dikotome opgaver?
2) DNT benytter flere forskellige opgavetyper defineret ved en kombination af en opgavetype og antal af items inden for opgaven. Spørgsmålet er, om der er visse opgavetyper, der giver mere sikre målinger af elevdygtigheden end andre, og om det er godt med få eller mange items i opgaven.
3) Vælger DNTs adaptive algoritme opgaverne på den bedst tænkelige måde?
4) Udover spørgsmålet om usikkerheden på målingerne er der et spørgsmål, om der er tilstrækkeligt med opgaver til ethvert niveau af elevdygtighederne. Hvis dette ikke er tilfældet, kan det vises at påvirke usikkerheden af resultaterne for de dygtigste og de svageste elever. Det er derfor også nødvendigt at spørge, om der er visse opgavetyper, der har problemer med at levere lette eller vanskelige opgaver, og om antallet af items i polytome opgaver har betydning for dette.

For at besvare disse spørgsmål er man nødt til, at beregne hvor megen information som svarene på de enkelte opgaver i givet fald kan bidrage med, og hvor meget denne information kan bidrage med til at forbedre sikkerheden.

Spørgsmål 1 er derfor et spørgsmål om, der er mere eller mindre information i polytome opgaver end i dikotome.

Spørgsmål 2 er et spørgsmål om, hvilke opgavetyper og – formater, der giver mest information.

Den adaptive algoritme skal altid forsøge, at vælge de opgaver der giver mest information, og spørgsmål 3 er derfor et spørgsmål om, hvorvidt det rent faktisk er tilfældet.

Baggrunden for spørgsmål 4 er, at hvis der ikke er tilstrækkelig mange lette og vanskelige opgaver, må den adaptive algoritme udtrække enten for nemme eller svære opgaver til eleven, og dermed vil svaret bidrage med relativ lidt information. Derfor er det værd at undersøge, om visse opgavetyper særligt egner sig til lette og svære opgaver.

Spørgsmålet om, hvordan man beregner den information, som svarene på en opgave kan bidrage med, og hvordan man kan beregne usikkerheden (SEM) ud fra informationen i de opgaver, der er besvaret, vil blive forklaret i afsnit 7.

## 4. Datagrundlaget

Ovenstående spørgsmål vil blive besvaret for 438 polytome og 672 dikotome opgaver fra profilområdet tekstforståelse i dansk, læsning i 4. klasse og for de tre profilområder i fysik/kemi i 8. klasse.

I disse profilområder anvendes der syv forskellige opgavetyper i forbindelse med de polytome opgaver, og antallet af items pr. opgave varierer fra 2 til 9. Opgavetyperne er (jf. tabel 1) imidlertid meget forskelligt fordelt på disse fire profilområder, og der er visse opgavetyper, der næsten ikke bruges. Det er påfaldende, at fysik/kemi er domineret af multiple choice opgaver med skemaformat. Der er derfor grund til at være særlig opmærksom på, om denne opgavetype vanskeliggør, at der opnås tilfredsstillende statistisk sikkerhed af målingerne af elevdygtighederne i fysik/kemi i 8. klasse.

**Tabel 1. Fordeling af polytome opgaver med hensyn til opgavetype og profilområde.**

| | Tekstforståelse Dansk, læsning 4. klasse | Energi Fysik/kemi 8. kl. | Fænomener mm. Fysik/kemi 8. kl. | Anvendelser mm. Fysik/kemi 8.kl. |
|---|---|---|---|---|
| Cloze-test | 41 | 6 | 5 | 3 |
| Indsættelse | | 13 | 10 | 6 |
| Match | | 1 | 9 | 9 |
| MC[1] skemaformat | 41 | 44 | 79 | 111 |
| MC[1] korrekte svar | | 17 | 10 | 18 |
| Hotspot | | 1 | 3 | |
| Indsæt i billede | | | 1 | |
| I alt | 82 | 92 | 124 | 146 |

[1]: Multiple choice

Udover opgavetypen og antallet af items pr. opgave vil usikkerheden i målingerne først og fremmest afhænge af, hvor mange opgaver der stilles, og hvor mange items eleven besvarer. Tabellerne 2 og 3 giver disse oplysninger for de fire profilområder i forbindelse med afviklingen af de nationale test i foråret 2015. I disse tabeller indgår alle opgaver, altså både polytome og dikotome.

Hvis man er bekymret over situationer, hvor usikkerheden er større end forventet, er det ikke nok at se på, hvor mange opgaver eleverne i gennemsnit kan nå at besvare. Man er nødt til at se på fordelingen af antal opgaver og items pr. elev og især på, hvor få opgaver, der i værste fald besvares. 5 % fraktilerne i tabel 2 og 3 giver et indtryk af dette. I værste fald må man forvente, at eleverne når at besvare ca. 10 opgaver i løbet af den tid, der er til rådighed. I forbindelse med tekstfor-

ståelse i 4. klasse vil det svare til ca. 14 items. I fysik/kemi, hvor der tilsyneladende bruges flere polytome opgaver end i tekstforståelse, er antallet af items i værste fald omkring 20 items.

**Tabel 2. Antallet af opgaver elever i gennemsnit nåede i hvert profilområde i foråret 2015**

| Test | Profilområde | 5 % fraktil | Gennemsnit | 95 % fraktil |
|---|---|---|---|---|
| Dansk, læsning 4. kl. | Tekstforståelse | 12 | 22 | 38 |
| Fysik/kemi 8. kl. | Energi | 10 | 20 | 33 |
| | Fænomener | 10 | 20 | 33 |
| | Anvendelser | 10 | 20 | 33 |

**Tabel 3. Antallet af items elever i gennemsnit nåede i hvert profilområde i foråret 2015**

| Test | Profilområde | 5 % fraktil | Gennemsnit | 95 % fraktil |
|---|---|---|---|---|
| Dansk, læsning 4. kl. | Tekstforståelse | 14 | 25 | 43 |
| Fysik/kemi 8. kl. | Energi | 20 | 39 | 66 |
| | Fænomener | 18 | 37 | 61 |
| | Anvendelser | 21 | 42 | 69 |

## 5. Om fejl i pædagogiske test

Der er to forskellige kilder til fejl i forbindelse med pædagogiske test: systematiske fejl og usystematiske fejl på grund af den statistiske usikkerhed, som altid vil være til stede i forbindelse med statistiske opgørelser og analyser.

Systematiske fejl i forbindelse med pædagogiske test kan forekomme, og de vil typisk forekomme:

1) hvis de psykometriske modeller og metoder, som anvendes i forbindelse med bearbejdningen af testresultaterne, er forkert specificeret. For eksempel fordi opgaverne afhænger af flere forskellige færdigheder eller fordi opgaverne ikke er lokalt uafhængige,
2) hvis sværhedsgraderne er forkerte,
3) hvis elevens testadfærd er afvigende i forhold til det, som de psykometriske modeller og metoder forventer.

Risikoen for de to første former for fejl kan minimeres, hvis afprøvningen af tilpasningen mellem opgaver og skalamodellen har været tilstrækkelig omhyggelig.

Systematiske fejl, der skyldes, at eleven af den ene eller anden grund ikke ”samarbejder”, kan aldrig undgås fuldstændigt, men den indledende afprøvning af testen kan afsløre, om det forekommer i et væsentligt omfang. For at undgå målefejl af denne årsag, kan man afprøve opgaverne

før de anvendes i egentlig test. Afprøvning afslører, at eleven svarer forkert (eller for så vidt korrekt) på opgaver, der burde være lette for eleven i forhold til det, som eleven har præsteret i starten af testforløbet, således at testresultatet ikke tages for gode varer. Dette afprøves allerede i forhold til DNT.

Denne rapport handler kun om usystematiske fejl på grund af den statistiske usikkerhed og om hvad, der kan gøres for at reducere usikkerheden til det mindst mulige. Det antages derfor 1) at tilpasningen mellem den Rasch-model, som anvendes i forbindelse med analyserne af DNT resultater, er blevet afprøvet, 2) at opgavernes sværhedsgrader er korrekt estimeret med så stor sikkerhed, at værdierne af sværhedsgraderne betragtes som kendte parametre, og 3) at testadfærd, der afviger fra det, som Rasch-modellen forventer, kun forekommer i meget begrænset omfang.

## 6. Om usikkerhed ifølge klassisk testteori

Psykometrien skelner mellem klassisk- og moderne testteori. Selvom udviklingen og anvendelsen af DNT er baseret på moderne testteori, vil det være nyttigt med et kort resume af den klassiske testteori. Mange af de centrale begreber, der dukker op i forbindelse med diskussionen af pædagogiske tests kvaliteter, er nemlig først defineret i den klassiske testteori.

Klassisk testteori betragter et testresultat, S som en funktion af en sand score, TS, og en fejl, E.

$$S = TS + E$$

Den klassiske testteori antager desuden, at fejlen er uafhængig af den sande score, og at den er normalfordelt med middelværdi lig med 0 en standardafvigelse, der omtales som ”standard error of measurement”, SEM. Det er denne størrelse, SEM, der er et udtryk for, hvor usikker testen er.

Inden for den klassiske testteori defineres desuden reliabiliteten, som forholdet mellem variansen af den sande score og variansen af den observerede score, som – fordi TS og E antages at være uafhængige – kan omskrives som en funktion af variansen af den sande score og SEM.

$$r = \frac{\mathrm{Var}(TS)}{\mathrm{Var}(S)} = \frac{\mathrm{Var}(TS)}{\mathrm{Var}(TS) + \mathrm{SEM}^2}$$

Grunden til, at reliabiliteten omtales her, er, at denne størrelse af mange opfattes som et udtryk for testens sikkerhed. Da reliabiliteten blandt andet afhænger af variansen af den sande score i en konkret elevpopulation, er denne opfattelse forkert. Reliabiliteten er et udtryk for, hvorledes testen fungerer i en konkret population, og tallet kan hverken generaliseres til andre populationer eller bruges som et udtryk for, hvor store forskelle man kan forvente at observere, hvis testen gentages, eller hvis der (som for eksempel i forbindelse med adaptive test) bruges andre opgaver, når testen gentages. Reliabiliteten fortæller, hvor god testen er til at sortere eleverne i den

konkrete population, fordi reliabiliteten er tæt på sandsynligheden for, at den dygtigste af to tilfældigt udvalgte elever får den bedste score. Men reliabiliteten er ikke et udtryk for, om målingen i sig selv er usikker. For at vurdere om det skulle være tilfældet, er det nødvendigt at beregne et tal for SEM og at lade det stå alene uden reference til spredningen af elever i den population, som man har undersøgt.

På trods af dette forbehold over for forståelsen af reliabiliteten indeholder klassisk testteori nogle tommelfingerregler, der knytter størrelsen af SEM til reliabiliteten. Ifølge en tommelfingerregel betragtes en SEM på 0,30 som acceptabel, fordi reliabiliteten vil være tæt på 0,90, hvis SEM er lig med 0,30 *og variansen af den sande score er lig med 1*[3].

Det er bevidst, at ovennævnte forudsætning om spredningen af eleverne er kursiveret, fordi forudsætningen ofte glemmes, når man diskuterer usikkerheden i forbindelse med pædagogiske test i almindelighed og DNT. Da der ikke er formuleret forudsætninger om, at testresultaterne fra DNT skal have en varians, der er lig med 1 i den elevpopulation, som testene er beregnet til, har et krav om, at SEM skal være lig med 0,30 ingen substantiel mening[4].

## 7. Om usikkerhed ifølge moderne testteori

Moderne testteori, der omtaler opgaverne som items, antager, at svarene på opgaverne kun afhænger systematisk af en bagvedliggende latent (ikke observerbar) variabel, der er et udtryk for den færdighed eller den egenskab, som testen forsøger at måle og af en række egenskaber ved opgaven.

For at beskrive den måde, som færdigheden påvirker svarene på opgaverne på, opstilles en statistisk model, der for hvert enkelt item angiver *sandsynligheden* for, at en elev med en dygtighed angivet ved en parameter, $\theta$, svarer korrekt på opgaven. Sådanne modeller omtales som Item response theory, IRT modeller. IRT modellen angiver altså den betingede sandsynlighed, $P(Item_i|\theta)$, for, at den i'te opgave besvares korrekt, hvis dygtigheden er lig med $\theta$.

Udover at det antages, at IRT modellen indeholder de korrekte matematiske funktioner, der definerer sandsynlighederne, antages det også, at svarene på de enkelte opgaver kun afhænger af færdigheden, og at svarene på én opgave ikke påvirker eller påvirkes af svarene på andre opgaver. Denne egenskab, som psykometrien omtaler som et krav om *lokal* uafhængighed, findes i alle gængse IRT modeller.

Værdien, $\theta$, optræder sammen med en række item-parametre, som en person-parameter i de funktioner, der definerer sandsynlighederne. Da dette $\theta$ er et udtryk for elevens færdighed, er det denne parameter, som man gerne vil vide noget om, når testresultatet gøres op. Og da det drejer sig om en ukendt parameter i en statistisk model, bruges et statistisk estimat af $\theta$ som et mål for,

[3] Jf. kapitel 7 i Wainer (2000).

[4] Hvis man endelig vil definere et krav svarene til kravet om SEM=0,30, når fordelingen af den sande scorer ikke er standardiseret, skal det være, at SEM = 0,30 $\times$ standardafvigelsen af den sande score i den aktuelle elevpopulation.

hvor dygtig eller mindre dygtig eleven er. For at skelne mellem på den ene side den sande parameter og estimatet af parameteren på den anden side, vil estimatet blive omtalt som $\hat{\theta}$ . $\theta$ svarer altså til det, der betragtes som den sande score i klassisk testteori, mens $\hat{\theta}$ svarer til den observerede score. Da der er tale om et statistisk estimat, er der også en statistisk standardfejl knyttet til estimatet, på samme måde som for alle andre statistiske estimater. Det er denne standardfejl, som moderne testteori opfatter som målingens SEM.

DNT bruger Rasch-modellen til at beregne sandsynlighederne for korrekte svar på opgaver, hvor der kun skelnes mellem rigtige og forkerte svar. Ifølge Rasch-modellen er en opgave karakteriseret ved en enkelt item-parameter, $\beta$, og sandsynligheden for et positivt svar på det i'te item afhænger af forskellen på personparameteren og parameteren for item'et:

$$P(\text{Item}_i \mid \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

Item-parameteren omtales som opgavens sværhedsgrad. Rasch-modellen påstår altså, at sandsynligheden for et korrekt svar er en funktion af forskellen mellem elevens dygtighed og opgavens sværhedsgrad, og at eleven har en 50 % chance for et rigtigt svar på en opgave, hvis dygtigheden er lig med sværhedsgraden. En opgaves sværhedsgrad er altså et udtryk for, hvor dygtig man skal være for at have 50 % chance for et rigtigt svar.

Udover antallet af opgaver afhænger SEM af både personparameteren og af sværhedsgraderne af de opgaver, som eleven har besvaret, uanset om der er tale om adaptive eller ikke-adaptive test. Beregningerne af SEM foregår i tre trin, hvor man starter med at beregne et mål for informationen fra de enkelte opgaver og slutter med at beregne SEM:

1) Først beregnes den såkaldte item-information, der er et kvantitativt mål for, hvor megen information svaret på en enkelt opgave indeholder om elevens dygtighed. Item-informationen er lav, hvis opgaven enten er alt for let eller alt for vanskelig for eleven, og den er størst, hvis sværhedsgraden er præcis den samme som dygtigheden.

2) Derefter beregnes test-informationen som summen af item-informationerne.

3) Til sidst kan SEM beregnes som $\text{SEM}(\theta) = 1/\sqrt{\text{Test inf}(\theta)}$

Hvis dikotome opgaver passer til en Rasch-model, er beregningen af item-informationen særlig enkel. Ifølge Rasch-modellen er item-informationen lig med:

$$\text{Item-information for item i} = \frac{\exp(\theta - \beta_i)}{\left(1 + \exp(\theta - \beta_i)\right)^2}$$

Item-informationen er altså en funktion af person-parameteren. Informationen går mod 0, når person-parameteren går imod $-\infty$ og $+\infty$. Den er størst, hvis person-parameteren er lig med item-

parameteren, hvor den altid er lig med 0,25, når der er tale om et dikotomt item fra en Rasch-model. Det følger heraf, at en dikotom opgave højst kan bidrage med en item-information, der er lig 0,25, og at den højeste testinformation der kan opnås, hvis der stilles dikotome opgaver, derfor er lig med k×0,25 svarende til en SEM lig med $2/\sqrt{k}$. Det forhold, at item-informationen går mod 0, jo dygtigere eller mindre dygtig eleven er, er et udtryk for, at opgaver, der er alt for lette eller alt for vanskelige for eleven ikke bidrager med noget, der kan gøre målingen af elevens færdighed mere sikker.

I almindelige ikke-adaptive test vil opgaverne have forskellige sværhedsgrader. Testinformationen vil derfor ikke være maksimal for nogen elever, og den vil for eksempel være beskeden for de meget dygtige elever, hvis testen indeholder meget få vanskelige opgaver og mange lette opgaver. Det er dette problem, der har motiveret udviklingen af adaptive test[5]. I stedet for at benytte opgaver, som vil bidrage med relativt lidt information for mange elever, forsøger adaptive test, at vælge opgaver, der i så høj grad som muligt bidrager med maksimal item-information for at opnå størst mulig testinformation og mindst mulig usikkerhed.

Tabel 4 viser den maksimale testinformation og den minimale SEM, som en adaptiv test med dikotome opgaver kan levere i forhold til antallet af opgaver, der besvares. Hvis den adaptive algoritme fungerer optimalt, og hvis item-banken indeholder tilstrækkelig mange lette og vanskelige opgaver, vil SEM ligge tæt på (men aldrig under) den usikkerhed, som man kan se i tabellen. Hvis der for eksempel stilles 20 dikotome opgaver, kan SEM aldrig blive mindre end 0,45. Tallet 0,45 kan derfor bruges som en benchmark værdi, hvis man både vil vurdere, hvor godt den adaptive algoritme har fungeret for en adaptiv test med 20 dikotome opgaver, og hvor godt en ikke-adaptiv test fungerer for elever med forskellige grader af dygtighed. Det kan for eksempel beregnes, at en ikke-adaptiv test med 20 opgaver, hvor sværhedsgraden er ligeligt fordelt fra -2,5 til +2,5, i bedste fald vil resultere i SEM = 0,54 og i værste fald (for meget dygtige og meget svage elever) med SEM = 0,82. Altså dårligere end en fungerende adaptiv test.

Eller med andre ord: Hvis den adaptive algoritme fungerer efter hensigten vil usikkerheden på elevdygtigheden i en adaptiv test altid være mindre end usikkerheden i almindelige ikke-adaptive test. Hvor meget mindre afhænger af opgavernes sværhedsgrader og af elevernes dygtighed.

[5] Wainer (2000) , Kapitel 1, ”Introduction and History”.

**Tabel 4. Optimal test-information og SEM i forhold til antal dikotome opgaver i adaptive test**

| Antal items | Test-information | SEM |
|---|---|---|
| 10 | 2,50 | 0,63 |
| 15 | 3,75 | 0,52 |
| 20 | 5,00 | 0,45 |
| 25 | 6,25 | 0,40 |
| 30 | 7,50 | 0,37 |
| 35 | 8,75 | 0,34 |
| 40 | 10,00 | 0,32 |
| 45 | 11,25 | 0,30 |
| 50 | 12,50 | 0,28 |

Som supplement til tabel 4 viser tabel 5, hvorledes usikkerheden ville være i en adaptiv test med dikotome items, hvis antallet af opgaver svarer til dem, som man kan se i tabel 3, og hvis den adaptive algoritme fungerer, som den skal.

**Tabel 5. Den mindste SEM, som en adaptiv test med dikotome opgaver kan opnå**

| Test | Profilområde | 5 % fraktil | Median | 95 % fraktil |
|---|---|---|---|---|
| Dansk læsning 4. kl. | Tekstforståelse | 0,53 | 0,40 | 0,30 |
| Fysik/kemi 8. kl. | Energi | 0,45 | 0,32 | 0,25 |
| | Fænomener | 0,47 | 0,33 | 0,26 |
| | Anvendelser | 0,44 | 0,31 | 0,24 |

## 8. Om usikkerheden i de nationale test

Beregningen af usikkerheden i de nationale test er kompliceret af forekomst af opgavetyper, som betyder, at forudsætningerne om lokal uafhængighed ikke kan være opfyldt, fordi der stilles flere forskellige delspørgsmål til samme emne. I forbindelse med afprøvningen af DNT, blev det derfor besluttet at betragte antallet af korrekte svar på delspørgsmål til et enkelt emne, som et super-item med flere forskellige svarmuligheder og at undersøge om disse super-items passede til Rasch-modeller for polytome items (såkaldte partial credit eller PCM modeller). I teorien for adaptive test, omtales sådanne opgaver som testlet-opgaver. Hvis afprøvningen af testlet-opgaver faldt heldigt ud, kunne målingen (estimationen af personparameteren) beregnes på samme måde som for Rasch-modeller for dikotome items, men beregningen af usikkerheden vil blive lidt mere kompliceret.

Matematisk set er modellen mere kompliceret end modellen med dikotome opgaver. I stedet for en enkelt sværhedsgrad afhænger antallet af korrekt besvarede delspørgsmål af en række opgave-

parametre omtalt som tærskelværdier. Disse kan ikke på nogen enkel måde sammenfattes som et udtryk for opgavens sværhedsgrad.

For at give et indtryk af PCM modellens kompleksitet viser tabel 6 sandsynlighederne for en model for en opgave med tre items, hvor man kan score 0, 1, 2 eller 3 point. Modellen afhænger af tre tærskel værdier, hvor $\beta_{ij}$, er den j'te tærskelværdi for den i'te opgave. Udover, at tabellen er inkluderet for at illustrere kompleksiteten i modellen, er hensigten også at gøre det klart, at der ikke findes samme enkle svar på, hvad det er, der karakteriserer opgavens sværhedsgrad, fordi opgaven er karakteriseret ved tre og ikke én parameter.

**Tabel 6. Partial credit model for super-item, der optæller antal rigtige svar på tre delspørgsmål**

| Antal rigtige på 3 delspørgsmål | Sandsynlighed |
|---|---|
| 0 | $\frac{1}{1+\exp(\theta-\beta_{i1})+\exp(2\theta-\beta_{i1}-\beta_{i2})+\exp(3\theta-\beta_{i1}-\beta_{i2}-\beta_{i3})}$ |
| 1 | $\frac{\exp(\theta-\beta_{i1})}{1+\exp(\theta-\beta_{i1})+\exp(2\theta-\beta_{i1}-\beta_{i2})+\exp(3\theta-\beta_{i1}-\beta_{i2}-\beta_{i3})}$ |
| 2 | $\frac{\exp(2\theta-\beta_{i1}-\beta_{i2})}{1+\exp(\theta-\beta_{i1})+\exp(2\theta-\beta_{i1}-\beta_{i2})+\exp(3\theta-\beta_{i1}-\beta_{i2}-\beta_{i3})}$ |
| 3 | $\frac{\exp(3\theta-\beta_{i1}-\beta_{i2}-\beta_{i3})}{1+\exp(\theta-\beta_{i1})+\exp(2\theta-\beta_{i1}-\beta_{i2})+\exp(3\theta-\beta_{i1}-\beta_{i2}-\beta_{i3})}$ |

Uanset kompleksiteten i modellen er det muligt at beregne opgave-informationen ud fra PCM modellen[6] og derefter at beregne test-informationen som summen af opgave-informationerne. I forbindelse med polytome opgaver defineres item-informationen herefter som opgave-information divideret med antallet af items i opgaven. Item-informationen i forbindelse med DNTs opgaver er altså defineret på en lidt anden måde end, når der normalt tales om pædagogiske test. For dikotome DNT opgaver er opgave-informationen dog lig med item-informationen i både den forstand, som der tales om her og i den forstand, som der normalt tales om i forbindelse med pædagogiske test.

[6] Item-informationen for items fra Rasch-modeller er – uanset, om der er tale om dikotome eller polytome items – lig med variansen af item-scoren givet værdien af personparameteren. Der henvises til Kreiner & Christensen (2013) for yderligere information om estimation af personparameteren i Rasch-modeller. Interesserede læsere kan finde de konkrete formler, der skal til for at beregne item-informationen for polytome opgaver med tre items i Appendiks 1.

I forbindelse med dikotome opgaver er det enkelt at udvikle en adaptiv algoritme, der vælger opgaver, der svarer til det aktuelle estimat af elevens dygtighed, fordi opgave-informationen er størst, hvis opgavens sværhedsgrad er lig med dygtigheden. Den adaptive algoritme skal derfor blot vælge en ny opgave blandt de opgaver, der har sværhedsgrader tæt ved det aktuelle estimat af dygtigheden.

I forbindelse med polytome opgaver er dette mindre enkelt, fordi definitionen af opgavens sværhedsgrad og relationen mellem tærskelværdierne, sværhedsgraderne og opgave-informationen er kompliceret.

I stedet for at kunne nøjes med et enkelt tal der karakteriserer opgaven, er der mindst tre forskellige indikatorer, der kan benyttes, når næste opgave skal vælges:

| | |
|---|---|
| Opgavens ”location” | = gennemsnittet af tærskelværdierne. |
| Opgavens ”difficulty” | = den dygtighed, der skal til for, at det i statistisk forstand forventede resultat på opgaven er lig med halvdelen af antal items. |
| Opgavens ”target” | = den dygtighed, hvor opgave-informationen er maksimeret. |

For dikotome opgaver er location, difficulty og target en og samme værdi. Det samme kan vises at være tilfældet for de fleste polytome opgaver med to items og for polytome opgaver, hvor tærskelværdierne ligger symmetrisk omkring opgavens location. Target-værdierne kan dog i ekstreme tilfælde ligge langt fra location og difficulty, hvilket kan føre til valg af informationsløse opgaver, hvis man bruger location eller difficulty som kriterie. I andre tilfælde kan der derimod være stor forskel på de tre værdier. For at illustrere dette viser tabel 7 tallene for to opgaver med fem items i tekstforståelse i dansk, læsning i 4. klasse.

For at kunne sammenligne den måde de to polytome DNT opgaver fungerer på, indeholder tabel 7 også en søjle med tærskelværdier og information fra en polytom opgave defineret som summen af fem dikotome opgaver med (næsten) samme target som de to polytome opgaver[7].

[7] Summen af et vist antal dikotome Rasch items kan vises altid at følge samme fordeling som et polytomt Rasch item (jf. Mesbah & Kreiner, 2013).

**Tabel 7. To opgaver med fem items i dansk, læsning i 4. klasse**

| Opgave | Opgave A | Opgave B | 5 dikotome items |
|---|---|---|---|
| Tærskel 1 | -0,61 | 1,10 | -1,96 |
| Tærskel 2 | -0,68 | -0,06 | -1,04 |
| Tærskel 3 | -0,26 | -0,48 | -0,35 |
| Tærskel 4 | 0,58 | -0,76 | 0,34 |
| Tærskel 5 | 1,72 | -1,51 | 1,26 |
| | | | |
| Location | 0,15 | -0,34 | -0,35 |
| Difficulty | -0,02 | -0,35 | -0,35 |
| Target | -0,33 | -0,36 | -0,35 |
| | | | |
| Opgave-information i target | 1,59 | 4,72 | 1,25 |
| Item-information | 0,32 | 0,94 | 0,25 |

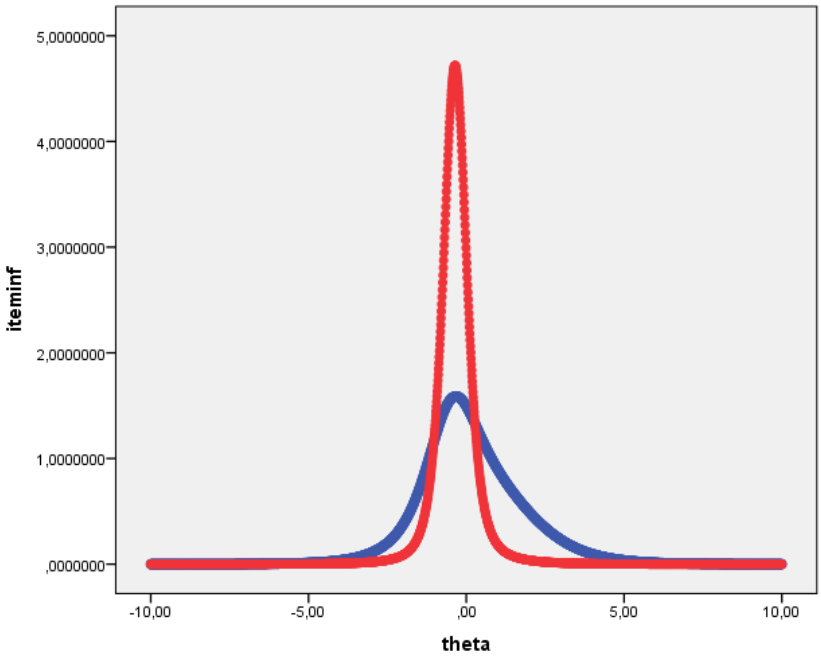
Opgave A og B er karakteriseret ved at maksimere opgave-informationen i næsten samme værdi (henholdsvis -0,33 og -0,36) af dygtigheden. Da formålet med den adaptive algoritme i DNT er at vælge opgaver, der giver så megen opgave-information dermed så lav SEM som muligt, vil disse opgaver være blandt kandidaterne, hvis dygtigheden ser ud til at ligge i nærheden af -0,35.

Konsekvenserne af valget er imidlertid meget forskellige for de to opgaver.

For det første kan det bemærkes, at sværhedsgraderne er forskellige. For den anden opgave (B) er difficulty næsten lig med target. En elev, som præsenteres for denne opgave vil derfor have en forventet opgave-score på 2,5. Sværhedsgraden for den første opgave (A) er derimod større. Hvis eleven i stedet for præsenteres for denne opgave, vil den forventede opgave-score være mindre en 2,5. Hvis man derfor mener, at det er et problem, at eleverne præsenteres for udfordrende opgaver med en stor risiko for fejl, er det klart, at den anden og lettere opgave vil være at foretrække frem for den første.

I forbindelse med ønsket om at estimere dygtigheden med så stor sikkerhed som muligt er sværhedsgraden et mindre problem i forhold til det, at der er meget stor forskel på den information, som de to opgaver kan levere. Den anden opgave med opgave-information lig med 4,72 giver næsten tre gange så megen information som den første, hvor opgave-informationen er lig med 1,59.

Problemet er illustreret i figur 1, der viser opgave-informationerne som funktioner af dygtigheden. Udover at vise den dramatiske forskel på opgave-informationerne for personer, der ligger i target for de to opgave, viser figuren, at opgave B (rød) kun fungerer bedst i intervallet fra -1,03 til 0,27. Mens opgave A (blå) til gengæld giver lidt bedre information for personer, der ligger længere fra target.




Figur 1. Opgave-information (omtalt som "iteminf" i figuren) for opgaverne A (blå) og B (rød) som funktion af dygtigheden

Udover at vise, at den ene polytome opgave er meget mere informativ end den anden, afslører tabel 7 også, at begge opgaver er mere informative end en polytom opgave bestående af fem uafhængige dikotome opgaver med sværhedsgrader, der svarer til target for de to DNT opgaver. Opgave-informationen for en sådan opgave er lig med 1,25, fordi hver af de fem dikotome opgaver bidrager med en item-information på 0,25 og fordi, alle fem opgaver havde den samme sværhedsgrad. Hvis sværhedsgraderne havde været forskellige, men med en gennemsnitsværdi tæt på -0,35, ville den samlede opgave-information være mindre end 1,25. Altså endnu dårligere end opgave A.

## 9. Analyser

Den adaptive algoritme i den aktuelle udgave af DNT vælger opgaver med target-værdier, der ligger tæt på det aktuelle estimat af dygtigheden, men uden hensyntagen til den opgave-information, som opgaverne leverer. Ovenstående opgave A og B vil derfor have næsten samme chance for at blive udvalgt, hvis det aktuelle estimat af dygtigheden er lig med -0,35, men udbyttet – i form af større sikkerhed – ville være meget mindre, hvis opgave A blev valgt i stedet for opgave B.

Det skal understreges, at den nuværende adaptive algoritme ikke kan betragtes som dårlig. Den leverer allerede pædagogiske test med høj statistisk sikkerhed. På trods af at det ikke er muligt at generalisere ud fra et enkelt eksempel med to opgaver, illustrerer eksemplet, at den adaptive algoritme måske kan forbedres, og at de indledende spørgsmål i afsnit 3 er relevante at få besvaret.

Derfor blev target-værdien, opgave-informationen og item-informationen beregnet for hver opgave, hvorefter opgaverne blev klassificeret i forhold til, hvor megen information opgaverne kunne give sammenlignet med den information, som kan opnås ud fra et tilsvarende dikotome opgaver.

Der kan skelnes mellem følgende fire grupper:

1) Opgaver med item-information der er mindre end 0,20, som derfor fungerer dårligere end et tilsvarende antal dikotome opgaver.

2) Opgaver med item-information fra 0,20 – 0,30 der fungerer på samme måde som et tilsvarende antal dikotome items.

3) Opgaver med item-information fra 0,31 – 0,50 der fungerer bedre end et tilsvarende antal dikotome opgaver.

4) Opgaver med item-information der er større end 0,50, og som derfor fungerer meget bedre end et tilsvarende antal dikotome opgaver.

Opdelingen af opgaver med hensyn til item-informationen set i forhold til item-informationen fra dikotome opgaver skulle kun tjene som et første forsøg på at skabe overblik over resultaterne. De præcise værdier af item-informationer og targets blev efterfølgende brugt til statistiske varians-analyser af effekten af opgavetype og antal items på de to forhold. Da opgave-informationen i target altid er lig med 0,25 for dikotome opgaver omfattede analysen af opgavetypernes betydning for informationen kun de polytome opgaver. Analysen af sværhedsgraderne omfattede til gengæld både dikotome og polytome opgaver.

Tabel 8-11 viser fordelingen af de polytome opgaver fordelt med hensyn til opgavetype og item-information, hvorefter afsnit 10 opsummerer resultaterne ved at svare på de spørgsmål, der blev stillet i afsnit 3 og ved at diskutere konsekvenserne af resultaterne for det fortsatte arbejde med at forsøge at forbedre sikkerheden i DNT.

**Tabel 8. Fordeling af polytome opgaver i tekstforståelse i dansk, læsning i forhold til opgavetype og item-information**

| | Item-information i forhold til item-information for dikotome opgaver | | |
|---|---|---|---|
| Opgavetype | Som dikotome | Bedre | Meget bedre |
| Cloze-test | 12 | 17 | 12 |
| Multiple choice (skema) | 1 | 33 | 7 |

**Tabel 9. Fordeling af polytome opgaver i energi og energiomsætning i forhold til opgavetype og item-information**

| | Item-information i forhold til item-information for dikotome opgaver | | | |
|---|---|---|---|---|
| Opgavetype | Dårligere | Som dikotome | Bedre | Meget bedre |
| Cloze-test | 0 | 3 | 3 | 0 |
| Indsættelsesopgave | 0 | 3 | 10 | 0 |
| Multiple choice (skema) | 0 | 15 | 36 | 3 |
| Multiple choice[1] (1+ svar) | 4 | 8 | 5 | 0 |
| Andet[2] | 0 | 0 | 1 | 1 |

[1]: En af disse opgaver har kun ét svar. De øvrige har flere svar.

[2]: Denne gruppe indeholder to forskellige typer: Hotspotopgave og matchopgave

**Tabel 10. Fordeling af polytome opgaver i energi og energiomsætning i forhold til opgavetype og item-information**

| | Item-information i forhold til item-information for dikotome opgaver | | | |
|---|---|---|---|---|
| Opgavetype | Dårligere | Som dikotome | Bedre | Meget bedre |
| Cloze-test | 1 | 1 | 3 | 0 |
| Indsættelsesopgave | 0 | 2 | 6 | 2 |
| Match | 0 | 0 | 7 | 2 |
| Multiple choice (skema) | 1 | 20 | 50 | 8 |
| Multiple choice (1+ svar) | 8 | 3 | 0 | 0 |
| Andet[1] | 0 | 0 | 2 | 2 |

[1]: Denne gruppe indeholder tre hotspotopgaver og en indsæt-i-billedopgave.

**Tabel 11. Fordeling af polytome opgaver i anvendelser og perspektiver i forhold til opgavetype og item-information**

| | Item-information i forhold til item-information for dikotome opgaver | | | |
|---|---|---|---|---|
| Opgavetype | Dårligere | Som dikotome | Bedre | Meget bedre |
| Cloze-test | 1 | 0 | 2 | 0 |
| Indsættelsesopgave | 0 | 2 | 1 | 3 |
| Match | 0 | 1 | 7 | 1 |
| Multiple choice (skema) | 1 | 30 | 66 | 14 |
| Multiple choice (1+ svar) | 7 | 10 | 1 | 0 |

## 10. Resume af resultater og konklusioner

De spørgsmål, der blev formuleret i afsnit 3, kan besvares på følgende måde.

*Giver DNTs polytome opgaver mere sikre eller usikre målinger end et tilsvarende antal dikotome opgaver?*

Svaret på dette spørgsmål er entydigt. DNTs polytome opgaver giver generelt den samme eller mere information end de dikotome opgaver. I mange tilfælde endda meget mere information end dikotome opgaver. Den ene undtagelse er for enkelte multiple choice opgaver i fysik/kemi, hvor der forekommer polytome opgaver, der er mindre informative end dikotome opgaver.

*Afhænger sikkerheden i testresultaterne af opgavetypen?*

Opgave-informationen fra opgaver i tekstforståelse i dansk, læsning i 4.klasse afhænger ikke af opgavetypen.

Det samme er tilfældet i fysik/kemi, bortset fra at multiple choice opgaver med ét eller flere svar giver mindre information end de andre opgavetyper. Og bortset fra, at der er enkelte sjældent brugte opgavetyper, der ser ud til at give mere item-information end andre typer. Disse forskelle er ikke statistisk signifikante, på grund af det lille antal opgaver af disse opgavetyper.

*Afhænger sikkerheden af testresultaterne af antallet af items i opgaverne?*

Her er svaret også entydigt. Opgaver med mange items bidrager med mere information pr. item end opgaver med få items.

Som konsekvens heraf vil opgavetyper, der lægger op til flere items end andre opgavetyper, være mere informative, selvom opgavetypen i sig selv ikke betyder noget.

*Vælger DNTs adaptive algoritme opgaver på den bedst tænkelige måde?*

Analyserne viser, at der kan være store forskelle på item-informationen blandt opgaver med samme opgavetype og samme antal items. Det tager DNTs adaptive algoritme ikke højde for. Det betyder, at sikkerheden vil kunne forbedres ved at fokusere på den information, som opgaverne kan levere, uanset om det sker der, hvor opgaverne giver mest information.

*Har opgavetype og antal items pr. opgave betydning for opgavernes sværhedsgrad?*

Ikke for opgaverne i tekstforståelse i dansk, læsning, men i høj grad for opgaverne i fysik/kemi, hvor multiple choice opgaver i skemaform giver relativt lette opgaver, og hvor antallet af items i flere tilfælde også bidrager til at reducere sværhedsgraden.

*Hvad kan man fremadrettet gøre for at forbedre sikkerheden i DNTs resultater?*

Det anbefales, at der i højere grad end tidligere satses på polytome opgaver frem for dikotome opgaver. Årsagen er, at polytome opgaver tidsmæssigt kan forventes at være mere effektive end dikotome, og fordi polytome opgaver i mange tilfælde bidrager med meget mere information pr. item end dikotome opgaver. Det anbefales også at satse på polytome opgaver med relativt mange items, fordi antallet af items også synes at have en positiv effekt på informationen pr. item.

Samtidig skal det bemærkes, at variationen i item-informationen for opgaver med samme opgavetype og samme antal items er så stor, at det anbefales, at der gennemføres indholdsanalyser af opgaverne af fagpersoner for at skabe bedre overblik over de faktorer, der bidrager til item-informationen, før der udvikles nye opgaver. Der er meget at hente med større klarhed over, hvad det er, der gør, at visse opgaver er mere informative end andre opgaver.

De polytome opgaver i fysik/kemi er domineret af multiple choice opgaver i skemaformat. Da disse opgaver erfaringsmæssigt er forholdsvis lette, er denne opgavetype mindre velegnet for de dygtigste elever ud fra en statistisk betragtning. Der er dog eksempler på opgavetyper (f.eks. hotspot-, indsættelses-, og matchopgaver), der bruges relativt sjældent, men som både er mere informative og vanskeligere. Selvom datagrundlaget vedrørende disse opgaver er for spinkelt til, at der formuleres håndfaste konklusioner, kan det alligevel anbefales, at der ses mere på disse opgaver og i givet fald fremover satses stærkere på disse opgavetyper.

Når alt dette er sagt, skal det samtidig understreges, at den adaptive algoritme bør ændres, hvis man skal drage det fulde udbytte af fordelene ved de polytome opgaver. Opgaverne udvælges altid ud fra det aktuelle estimat af elevens dygtighed. Den adaptive algoritme vælger i øjeblikket den næste opgave ud fra de opgaver, der sigter på elever med den estimerede dygtighed (opgavens target-værdi), men uden hensyntagen til hvor megen information opgaven i givet fald kan bidrage med. Hvis formålet udelukkende er, at minimere den statistiske usikkerhed på den beregnede elevdygtighed, kan den adaptive algoritme modificeres således, at algoritmen identificerer de opgaver, der giver mest item-information, der hvor eleven befinder sig, uanset om opgaven sigter skævt i forhold til eleven. Det afgørende skal kun være, om der er andre opgaver, der ville give endnu mere information, der hvor eleven befinder sig. Hvis det ikke er tilfældet skal opgaven kunne vælges. Der er flere måder, en sådan adaptiv algoritme kan implementeres. Den enkleste er en gang for alle at udarbejde en tabel, der for et vist antal af værdier på personskalaen indeholder en prioriteret liste over opgaverne med de opgaver, der bidrager mest først og dem, der bidrager mindst sidst.

Det eneste forbehold, som man kan have over for de polytome opgaver, er, at de kun leverer maksimal information i et måske snævert interval. Da elevens dygtighed er meget usikkert bestemt i starten af testforløbet, bør man fastholde, at der i starten af testforløbet kun bruges dikotome opgaver, sådan som det sker i øjeblikket, og at man først skifter til polytome opgaver, når man har et første bud på, hvor eleven befinder sig. I den sammenhæng er det vigtigt at minde om, at den måde den adaptive algoritme fungerer på i starten af testforløbet ikke er optimal, fordi processen starter uden nogen som helst oplysninger om, hvor man kan forvente at finde eleven. Hvis man vil reducere antallet af opgaver, der reelt ikke bidrager til at forbedre sikkerheden, kan man starte testforløbet med et kvalificeret gæt på, om der er tale om en dygtig eller mindre dygtig elev. Input til en sådan start kunne ideelt være lærerens forventninger til elevens præstationer eller oplysninger om, hvor eleven lå i forbindelse med tidligere testforløb. En sådan start vil måske ikke bidrage med meget for flertallet af eleverne, men det vil have betydning for sikkerheden i resultaterne for de stærkeste og de svageste elever.

## Referencer


Kreiner, S. & Christensen, C.B. (2013): Person Parameter Estimation and measurement in Rasch Models. I Christensen CB, Kreiner S, Mesbah M (red). *Rasch models in Health*. London: ISTE & Wiley and Sons, side 63-78.

Mesbah, M. & Kreiner, S. (2013): Rasch Models for Ordered Polytomous Items. I Christensen CB, Kreiner S, Mesbah M (red). *Rasch models in Health*. London: ISTE & Wiley and Sons, side 27-42

Wainer, H. (2000) *Computerized Adaptive Testing: A primer*. Second Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202

## Appendiks 1: Beregning af opgave-informationen i polytome opgaver

Antag at opgave i indeholder k forskellige dikotome items, og at den samlede score, $Y_i$ består af antallet af korrekte svar på disse items.

Fordelingen af $Y_i$ afhænger af elevens dygtighed repræsenteret ved personparameteren, $\theta$, og k såkaldte tærskelværdier, $\beta_{i1},\ldots, \beta_{ik}$, således at sandsynligheden, $p_{iy}(\theta)$, for at svare rigtigt på y ud af de k opgaver er lig med:

$$\mathbf{p_{iy}(\theta) = \frac{exp\left(y\theta - \sum_{j=1}^{y}\beta_{ij}\right)}{\sum_{x=0}^{k} exp\left(x\theta - \sum_{j=1}^{x}\beta_{ij}\right)}}$$

Opgave-informationen for polytome Rasch opgaver er lig med variansen af $Y_i$. For at beregne den er det derfor nødvendigt først at beregne elevens forvente opgavescore:

$$\mathbf{E(Y_i \mid \theta) = \sum_{y=0}^{k} y \cdot p_{iy}(\theta)}$$

og den forventede kvadrerede score:

$$\mathbf{E(Y_i^2 \mid \theta) = \sum_{y=0}^{k} y^2 \cdot p_{iy}(\theta)}$$

hvorefter opgave-informationen beregnes som:

$$\mathbf{OpgaveInf = E(Y_i^2 \mid \theta) - \left(E(Y_i \mid \theta)\right)^2}$$