# Test til måling af progression

## - Om at udvikle test til måling med ipsativ reference

Jakob Wandall, NordicMetrics & University of Aarhus, Dep. of Education

Siri Jordahn, NordicMetrics

***Keywords: testudvikling; progression; skalaækvivalering; kvasi-eksperiment; folkeskolereform***

*Artiklens hovedformål er at redegøre for metodeovervejelser og proces i forbindelse med udvikling af nye test. Baggrunden for udviklingen af de nye test er DRs dokumentarserie i skoleåret 2014/15 om folkeskolereformen, hvori elevernes faglige udvikling det første år efter implementeringen undersøges. DR kunne ikke få adgang til de nationale test (DNT) og bad derfor NordicMetrics om at udvikle test til formålet. I artiklen illustreres det, hvordan disse test (NM-test) er udviklet, kvalitetssikret, og det er belyst, hvordan det sikres, at de måler de samme latente egenskaber som DNT, samt hvordan de adskiller sig fra disse. NM-testene består af en præ- og posttest i fagene læsning (dansk), matematik og engelsk, som er udviklet på samme skala, så de kan vise elevers faglige progression henover et skoleår. Betingelser for retvisende måling af elevers faglige progression fremfor blot fagligt niveau beskrives i starten af artiklen, da dette fokus er det grundlæggende nye perspektiv for disse test. Det diskuteres til sidst, hvordan elevernes faglige progression i skoleåret 2014/15 kan sammenlignes med elevernes faglige progression før folkeskolereformen; samt om man som følge af reformen vil kunne forvente en udvikling i elevernes faglige progression det første år efter implementeringen. Prætestene blev gennemført i august 2014, mens posttestene gennemføres i juni 2015.*

## Abstract

This paper examines three themes, which have in common the shift in focus from measuring the absolute academic level (whether it is based on norm or criterion reference) towards a learning paradigm (individual student's growth). This is known as "ipsative reference", where it is the individual student's own academic level that constitutes the personal reference.

Firstly, well-functioning American items can provide a foundation for development of tests in reading (Danish), mathematics (translation and reformulation) and English in Danish schools. We described how the originally American items constitute the basis for the construction of three sets of pre- and post-tests in the subject areas reading (Danish), English and mathematics. All of the items in each subject area are tested systematically on Danish students. The responses are Rasch-analysed and scales have been formed for each of the three subject areas. In this process, items that do not fit the scales have been discarded.

Secondly, the article describes how these tests are constructed and how they match (through CCT equating) the national test in Denmark (DNT). The test construction process is described (Case: Mathematics). We show how these tests can be used to measure the academic level and growth (progression) on scales corresponding to those used in the Rasch model behind DNT. Each of the three set of items, which form the basis of a pre- and posttest, are scored on scales closely related to the scales from DNT. The items are tested in 5 schools in the Municipality of Aalborg, and the

results are presented on a national television channel (DR1) as a documentary in which students' progress during the first year of the new school reform. DR1 follows one class from 8th grade which could affect their performance. The other students on the five schools are used as a control group.

Thirdly, the paper discusses measurement of the effect of the Danish school reform. Is it likely to trace effects of the involved schools' implementation of the new school reform after one year? Considerations on measurement of student's progress are made and a method on how to compare before and after the reform is discussed. When the post-tests are administered in the summer of 2015, the data will form the basis of a study of the academic growth of the pupils involved. Although the scales are somewhat different in DNT and NM-tests, it might be possible to compare them, provided that both tests work with ipsative reference - in this method each pupil's progress rather than the absolute academic level is assessed. By comparing with previously obtained data from DNT on progress we intend when the posttests in june 2015 have been administered to investigate if there is a significant difference change in the academic progress before and after the reform. Finally the issues of causality are discussed.

# Indledning

### *Baggrunden for projektet*

Danmarks Radio og Aalborg kommune aftalte i foråret 2014 et samarbejde om en række dokumentarudsendelser om effekten af folkeskolereformen. Mere specifikt var DR interesseret i at undersøge elevernes faglige progression det første skoleår efter reformens implementering med udgangspunkt en 8.klasse på Sofiendalsskolen. Da der er opstillet kriterier for elevernes faglige mål i forbindelse med reformen ud fra DNT, blev det undersøgt, hvorvidt DR kunne få tilladelse til at anvende disse test til projektet. Imidlertid er resultater fra DNT fortrolige, og da målet med dette projekt netop var at kaste lys over faglige udvikling, bl.a. ved at belyse resultater med skolens, elevers og forældres billigelse, kunne DNT ikke anvendes i denne sammenhæng.

På det grundlag blev NordicMetrics[1] kontaktet med henblik på at udvikle og afprøve nye test, som kunne bruges til formålet. Der blev indledt et samarbejde mellem DR, Aalborg kommune og NordicMetrics om at udvikle en række test til formålet: At måle faglig udvikling på et videnskabeligt grundlag inden for læsning (dansk), matematik og engelsk. Yderligere skulle det sikres, at der måles nogenlunde tilsvarende egenskaber, som hvis man havde anvendt DNT. Der er trods ligheder mellem skalaerne fra DNT og NM-testene ikke tale om et parallelt system – men det er undersøgt, at NM-testene korrelerer i rimeligt omfang og således afdækker tilsvarende underliggende egenskaber som DNT i de tre fag. Kravet til NM-testene var, at de skulle give mulighed for ipsative vurderinger, altså at elevernes individuelle fremskridt blev vurderet uafhængigt af elevens faglige niveau. DR-dokumentarserien følger én 8. klasse på Sofiendalskolen, men alle 8.klasser på den pågældende skole samt 8. årgange på fire andre af Aalborgs skoler indgår i projektet, da de fungerer som kontrolgrupper. Dokumentarserien bygger altså på et kvasi-eksperiment med 13 deltagende klasser fordelt på fem skoler.

***Fagligt niveau eller progression?***

Der har siden 1993 været et stadigt stigende fokus på evaluering af elevernes faglige niveau og udvikling i den politiske debat om folkeskolen. DNT blev indført ved lov i 2005/06, og siden skoleåret 2009/10 har der været gennemført måling af niveauet på en række fagområder med de DNT.

Præstationer i DNT i læsning (dansk) og matematik er grundlaget for tre ud af de fire succeskriterier i forbindelse med den nye folkeskolereform; her er målet, at A) 80 % af eleverne skal være gode til at læse og regne målt med testene, at B) andelen af de allerdygtigste stiger år til år i dansk og matematik, samt at C) andelen af dårlige præsterede elever skal reduceres fra år til år uanset social baggrund (Undervisningsministeriet, 2014a). For alle tre måls vedkommende drejer det sig om, hvordan de enkelte elevårgange præsterer og sammenligning af resultaterne med tidligere årganges resultater (fx hvordan dette års elever i 6. klasse præsterer i DNT i forhold til hvordan eleverne, som gik i 6. klasse sidste år præsterede i samme test). Der er altså fokus på absolutte niveauer (målt i forhold til en norm eller faglige kriterier), som er fastsat i forbindelse med DNT, snarer end hvordan de enkelte elever på den enkelte årgang udvikler deres færdigheder hen over årene.

***Sammenligning af årgangsresultater og betydningen af elevens baggrund***

Når man anvender årgangsresultater som mål for kvaliteten af skolernes indsats, uanset om der arbejdes med norm- eller kriteriebaserede scorer, antages implicit, at elevgrundlaget er ens og direkte sammenligneligt fra år til år. Dette er generelt et meget usikkert grundlag at bygge på (Wainer, 2011). Ser man på PISA-resultater i læsning for Danmark i de 12 år, hvor Danmark har deltaget i alle 5 undersøgelser, ligger vi stabilt i læsning med mellem 490 og 500 PISA-point som gennemsnit. Det kunne tyde på, at et jævnt niveau over årene kunne være en rimelig antagelse på landsplan. På den anden side er der større udsving i både naturfag og matematik, hvilket peger i den modsatte retning (Egelund, 2013).

Yderligere spiller populationens størrelse en betydelig rolle. Jo færre elever i populationen des mindre holdbar bliver antagelsen om identiske årgange, der altså er tvivlsom på landsplan. Den bliver mere problematisk på kommuneniveau og især i mindre kommuner. Endnu større bliver problemer med sammenligning af årgange på skoleniveau, da det ofte ses, at skoler klarer sig godt i enkelte år, fordi den pågældende årgang indeholder nogle generelt meget dygtige elever (Rangvid, 2008). Statistisk korrektion for social baggrund kan kun i begrænset omfang rette op herpå, da de faktorer, som man kan få oplyst fra Danmarks Statistik (etnicitet, køn, forældres uddannelse, løn, stilling, familieforhold mv.), erfaringsmæssigt kun forklarer op til ¼ af variationen på elevniveau (Houlberg, 2013). Således har skoler, der hovedsageligt rekrutterer elever fra socialt belastede områder, gennemgående laverer gennemsnitsresultater, end skoler der rekrutterer fra villakvarterer med overvejende veluddannede beboere også efter social korrektion. Dette er ikke på grund af undervisningens kvalitet, men på grund af elevernes baggrund i bred forstand – fx gener, omgivelser, adfærd, omsorg og andre påvirkninger i de tidlige år (Hattie, 2005; Haertel, 2013).

Desuden: Des større spredning der er i elevernes forudsætninger ved skolestart, og des mindre den årlige progression er i forhold til spredning i det faglige niveau, jo mere følsomt bliver

gennemsnittet overfor enkelt elevers præstationer. Der skal ikke meget til at rykke ved årgangsgennemsnittet på en skole med eksempelvist ét eller to spor, hvor afstanden mellem top og bund svarer til 10 års progression, hvilket er ganske sædvanligt[2].

***Ipsativ tilgang: Fra absolut fagligt niveau til faglig progression***

Det er derfor ikke statistisk meningsfuldt at bedømme en skoles kvalitet ud fra dens elevers resultater i DNT eller folkeskolens afgangsprøve. Den læsefærdighed, en elev fx kan opnå i slutningen af skoleforløbet, afhænger i altovervejende grad af, hvilke sproglige forudsætninger eleven havde ved skolestart (Haertel, 2013). Når elever i slutningen af folkeskolen derfor ligger på meget forskellige niveauer, har lærerne blot indfriet den opgave, som der er stillet dem i folkeskolelovens § 18. Heri står der, at skolen skal planlægge og tilrettelægge undervisningen: *"(…) så den rummer udfordringer for alle elever."* og sørge for at: *"(…) den svarer til den enkelte elevs behov og forudsætninger."* (Undervisningsministeriet, 2014b). Konsekvensen at gøre dette effektivt er, at forskelle i det faglige niveau mellem stærke og svage elever opretholdes helt eller delvist. I stedet for at måle elevernes absolutte faglige niveau (uanset om det er op imod en norm- eller kriteriebaseret reference) der ikke tager hensyn til forskelle i skolers elevgrundlag, kan der arbejdes med ipsativ reference.

**Faktaboks: Forskellige former for referencer for test set i forhold til DNT**

**Normbaseret reference** betyder, at elevers resultater holdes op mod en norm, altså sammenlignes med en fordeling som afspejler en gruppe, der anvendes som reference (fx landsfordelingen af karakterer eller testresultater). Ved en ægte normbasering opdateres normen jævnligt (fx årligt), således at gennemsnittet/normen, der refereres til, altid afspejler det aktuelle niveau.

*DNTs skalaer (1-100 og 5-trinsskalaer) referer til den gruppe elever (ca. 14.000), der blev anvendt ved kalibreringen af testene. De er således ikke ægte normrefererende. Her er formålet netop, at kunne følge udvikling og normen ændres derfor ikke, selvom middelpræstationen har flyttet sig (principielt kan man ende op i en situation, hvor alle elever ligger over normgennemsnittet).*

**Kriteriebaseret reference** henviser til, at elevernes resultater holdes op imod kriterier, dvs. typisk kvalitative faglige mål, der er fastsat af eksperter/fagfolk. Men begrebet "kriteriebaseret" er ikke et entydigt defineret og velafgrænset begreb og i forskningslitteraturen på området findes et utal af forståelser af begrebet. Begrebet "criteria based" behandles således ofte som synonymt med "standard referenced", hvor fx kravet til at bestå målt ved en andel rigtige (fx procentscoring), betegnes som kriteriet (Sadler 2005).

*I DNT anvendes normen ved testsystemets udgangsår (dvs. 2009/10) som systemets standard. Denne kombinerede anvendelse af begreberne norm og standard (eller kriterie i bred forstand) indebærer formidlingsmæssige udfordringer. Kriteriebaseringen af de nationale test som gennemføres ultimo 2014 vil antageligt rette op på en del misforståelser.*

**Ipsativ reference** (kaldes også for selv-reference) anvendes, hvor fokus er på at sammenligne med egne individuelle tidligere resultater ("ipse-" er latin for "selv-"). Ved den klassiske anvendelse i den psykologiske teori, fx ved personlighedstestning, har alle respondenter ofte samme samlede score; hvor fokus er den enkeltes styrker og svagheder på delområder. Som udgangspunkt er scorer opgjort på ordinalniveau og er ikke direkte sammenlignelige mellem personer. (Dudley & Swaffield, 2008)

*Men man kan også anvende principperne med ipsativ reference for arbejde med individuel udvikling (progression el. "growth"), der registreres på norm- eller kriterieskalaer. Her vil sammenlignelighed være afgjort af skalaens egenskaber. Kun hvis scoren er opgjort på en intervalskala (fx som den bagved liggende Rasch-skala i DNT), vil man kunne sammenligne direkte.*

Pointen er, at når skolerne har meget forskellige elevgrundlag, så kan de også kun i begrænset omfang hæfte for slutresultater fx målt ved afgangsprøverne eller DNT - men skolerne og lærerne skal i betydeligt omfang kunne stå på mål for elevernes faglige udvikling (progression).

## Metodeovervejelser bag konstruktion af NM-test

For at kunne undersøge elevernes faglige progression over skoleåret 2014/15 er en egnet og ofte anvendt fremgangsmåde at udvikle en præ- og en posttest til anvendelse ved start og slut på undersøgelsesperioden (jf. fx Bonde, 2014). Testene skal, for hvert af de tre fagområder, kunne måle dygtighed på identiske intervalskalaer, så de to målinger (præ- og posttest) er direkte sammenlignelige. Hver test skal bestå af så tilpas mange items, at testene kan afvikles på så kort tid, at eleverne ikke mister koncentration, men så mange at man samtidigt opnår en rimelig nøjagtighed i målingen. Det har resulteret i, at hver NM-test omfatter 45-50 items (med et begrænset overlap på 5-10 items som optræder både i præ- og posttest). Det vil sige, at der er behov for i alt 90-95 afprøvede velfungerende items pr. fagområde.

Der er lånt amerikansk udviklede items af typen "embedded sentence end" til måling af engelsk[3]. Til brug for måling af læsning (dansk) er der udviklet danske items til dels efter samme skabelon som de amerikanske. Og endelig er der blev konstrueret items til måling af matematik med stærk inspiration fra det amerikanske items fra NAEP[4].

Alle disse items er tidligere afprøvet på amerikansk grund og har fungeret fornuftigt, men det er langtfra givet, at items kan overleve den kulturelle og sproglige rejse fra amerikanske elever til elever i den danske folkeskole (Wainer, 2011). Derfor blev de alle afprøvet i juni 2014 på ca. 200 elever på fem Aalborgskoler.

### *Afprøvning af NM-items og ækvivalering med resultaterne med DNT*

Følgende klassetrin har afprøvet opgaver i:

- 6. klasse i matematik
- 7. klasse i engelsk
- 8 klasse læsning (dansk)

Disse klassetrin er blevet udvalgt til afprøvning af items fordi, eleverne nogenlunde samtidigt (inden for 3 måneder) har anvendt DNT i samme fag, og at det derfor bliver umiddelbart meningsfuldt at ækvivalere resultaterne – trods den tidsmæssige distance.

Der vil erfaringsmæssigt være items, der ikke passer til skalaen og derfor må kasseres. For at have et overskud at kassere fra blev der afprøvet lidt flere items, end der var brug for.

Ambitionen har været at komme tættest muligt på tilpasning til en Rasch-model, givet at der skulle udvælges 80-95 items til hver test. På grundlag af elevbesvarelser af NM-testens items gennemføres en Raschanalyse, hvorved der blandt andet beregnes en dygtighed for hver elev – dygtigheden (også benævnt personparameteren) opgøres på en Raschskala, der hvor enheden er logitter. Elevdygtighederne på NM-testen kan sammenholdes med resultaterne fra DNT, hvor resultaterne for hvert profilområde ligeledes er målt i Rasch-logitter. For læsning er det som NM-testene måler svarende til det, der måles i profilområdet tekstforståelse i DNT i læsning (dansk).

En særlig problemstilling gør sig gældende for matematik og engelsk, hvor den egenskab som NM-testen måler ikke svarer til et specifikt profilområde, men til det samlede produkt af hele DNT. Den samlede percentilscore er nemlig dannet som et simpelt gennemsnit af de tre profilområders percentilscorer. Det betyder, at der ikke eksisterer en underliggende Raschscore, der kan bruges som sammenligningsgrundlag. Den pragmatiske løsning på dette problem er, at foretage en beregning af hvor meget hvert profilområde kan antages at have bidraget til den samlede score. Dette gøres ved at forklare NM-testens personparameterestimat med DNT logits for alle tre profilområder i en lineær regressionsmodel (metoden er nærmere beskrevet i afsnittet om konstruktion af NM-matematiktest).

***Kriterier for sammensætning af NM-test***

Items er blevet udvalgt efter følgende kriterier i den nævnte orden:

- Først udvælges items, der passer bedst til en Rasch-modellen, det vil sige har et fit-residual mellem -2,5 og 2,5 (Pallant, 2007, s.5).
- Dernæst udvælges items, som overdiskriminerer mellem eleverne, forudsat disse items korrelerer positivt på et rimeligt niveau med elevens score i DNT (mellem 0,6 og 0,85)
- Efterfølgende bliver tjekket, at alle items korrelerede positive på et acceptabelt niveau med scoren fra DNT (korrelationen skal være mindst 0,2).

Svagt diskriminerende items, og items der korrelerer lavt med DNT bliver altså frasorteret på grundlag af afprøvningen. Der gennemføres herefter en Raschanalyse hvor disse item indgår og der dannes en skala, som både person- og itemparameteren (sværheden) kan placeres på. Endvidere bliver det ved at korrelere testresultater for eleverne i NM-testene og DNT undersøgt, om denne skala passer nogenlunde med skalaer, som anvendes i DNT.

Det skal understreges, at det ikke bliver helt det samme, der måles i NM-testene og DNT: Dels er der ikke – som ved DNT – fuldstændig tilpasning til en Rasch-model, dels er opgaverne opbygget og administreret forskelligt og dels er der tale om test i fast format, hvor DNT er adaptive[5]. Dette giver en forskel, men det har heller ikke været ambitionen at opbygge et parallelt system til DNT. Hvis det var helt det samme, ville offentliggørelse af resultaterne også kunne anses for problematisk. I NM-testene vil en rangordning af eleverne derfor blive anderledes end i DNT. Det er imidlertid hensigten, at der skal være stærke ligheder, idet det er de samme konstrukter, som måles: Korrelationer mellem 0,6 og 0,85 vil være det forventede niveau - svarende til korrelationen mellem velfungerende test på et givent fagligt område (Hoover & Gough, 1990; Williams, 2010; Stenner, 1997).

## Resultater fra arbejdet med testkonstruktionen

Etablering af præ- og posttest i matematik, læsning (dansk) og engelsk er foregået efter samme metode, og denne vil blive eksemplificeret ved en nøjere gennemgang af resultaterne fra matematik. Fælles er, at:

- Der afprøves godt 100 items pr. fagområde på ca. 200 elever.
- Ved en Rasch-analyse identificeres de items, der passer dårligt til skalaen og de items, der diskriminerer dårligt elimineres, og det tjekkes, at der er en fornuftig korrelation mellem DNT og henholdsvis NM-testens elevresultater og items.

- På samme måde som ved matematik beregnes ved regression en samlet DNT-score for engelsk.
- For læsning (dansk) er det enklere, da profilområdet Tekstforståelse (den samlede læseforståelse) svarer til den samlede score, der kommer ud af NM-testene.

Nøgletal for præ- og posttest for alle tre fagområder findes i tabel 1.

| Tabel 1 | Matematik | Læsning | Engelsk |
|---|---|---|---|
| Antal items afprøvet | 112 | 108 | 89 |
| Antal items accepteret (diskriminerer mindst lige så godt som Rasch-items og korrelerer stærkt med DNT-total score)<br>- heraf med -2,5<FitResid<2,5 | 94<br>(93) | 92<br>(84) | 79<br>(64) |
| Antal Items i præ- og posttesten | 50/50 | 50/50 | 46/46 |
| Antal elever (heraf med DNT) | 206 (175) | 217 (205) | 258 (251) |
| PSI præ/post | 0,836/0,818 | 0,820/0,846 | 0,889/890 |
| Korrelation mellem NM og DNT | r = 0,74 | r = 0,75 | r = 0,87 |
| Korrelation mellem NM-præ- og posttest | r = 0,896 | r = 0,789 | r = 0,938 |

***Case: Konstruktion af NM-test i matematik***

Udviklingen af NM-testene beskrives med udgangspunkt i NM-matematiktesten.

NM-matematiktesten er dannet på grundlag af 112 items skrevet på dansk med stærk inspiration fra NAEP-items, som er offentligt tilgængelige (nogle er bare oversat andre er tilpasset danske forhold). Disse NM-items i matematik er kategoriseret i fem sub-skalaer efter deres "moderitems" i NAEP: a) Numbers, b) Algebra, c) Geometry, d) Statistics og e) Measurement. Opgavernes indhold og deres formater minder endvidere ud fra en overfladisk gennemgang ganske meget om de opgaver, som optræder i DNT.

På det grundlag blev de efterkategoriseret til profilområderne i DNT:

1. a) & b): Tal & Algebra (38 items)
2. c): Geometri (34 items)
3. d) & e): Matematik i anvendelse (40 items)

Præ- og postmatematiktestens items er afprøvet på 206 elever på 6. klassetrin i juni 2014, og langt de fleste elever har besvaret alle items. Der er dannet tre skalaer (en for hvert af de tre profilområder: tal & algebra, geometri og matematik i anvendelse) samt en samlet skala med 112 items. Rasch-analysen viser, at der ikke er fuldstændig tilpasning til en Rasch-model; hvilket sandsynlighedsvis hænger sammen med, at enkelte items, som overdiskriminerer, er bibeholdt (jf. s.6). Dette ses i figur 1, hvor fordelingen af elevernes dygtighed (de røde) er vist mod items sværhed (de blå); da items skal anvendes på 8. klassetrin, forventes det at være en fordel, at items er svære.

Figur 1

Der er identificeret 175 elever, som både har gennemført DNT og NM-testen på 6.klassetrin. For at kunne sammenligne skalaerne fra de DNT og NM-skalaerne må DNTs percentilresultater oversættes tilbage til de underliggende resultater (logits) på Rasch-skalaer. På grundlag af oplysninger fra UNI-C kan sammenhængen mellem percentil- og Rasch-scorer (som følger sigmuid-funktioner) for profilområderne i matematiktesten til 6. klasse beskrives. Disse sammenhænge er illustreret i figur 2.

Figur 2

En samlet Rasch-score for matematik kan ikke findes, da den samlede score kun eksisterer for percentilværdier (den er dannet som et almindeligt gennemsnit at de tre percentilscorer). Det er imidlertid hovedformålet at finde en samlet matematikscore. Rasch-scorerne for NM-testene (hovedskalaen og de tre underskalaer) og DNT Rasch-scorerne for de tre profilområder er alle transformeret til z-scorer[6].

En samlet score for DNT burde have taget udgangspunkt i en item-baseret estimation af alle responserne i de tre profilområder, men eftersom percentilscoren som nævnt ovenfor er det rå gennemsnit af profilområdernes percentilscorer, kan det ikke oversættes til en Rasch-score. Som en næstbedst løsning er det samlede resultat for NM-testen søgt forklaret i en lineær regressionsanalyse med z-scorer for de tre DNT-profilområder. Regressionslinjen kan benyttes som det bedst tilgængelige bud på et samlet resultat af DNT i matematik i 6. klasse. Det ovenstående beror på en brutto-opgørelse på grundlag af alle 112 items, og der er ikke heri taget stilling til, hvor godt items passer til en fælles NM-skala og/eller til DNTs skalaer.

Det er derfor undersøgt, om der er items, som passer dårligt til skalaen. Dette er gjort dels ved at se på fit-residualer til de enkelte items og dels ved at se på korrelationen mellem det enkelte item og resultater fra hhv. NM-testen og den beregnede samlede DNT-skala. Er fit-residualet stort, betyder det, at der er lille eller ingen sammenhæng mellem den enkelte elevs dygtighed og elevernes besvarelse af dette item.

Det viste sig, at disse parametre for tilpasning til modellen udpegede de samme items som gode og mindre egnede, og der blev på grundlag heraf frasorteret 18 items.

Et eksempel på en opgave med et højt fit-residualet, der er frasorteret (item 106 har et fitresidual på over 6) lyder: ”Hver af de 6 overflader af en terning er mærket enten R eller S. Når terningen er kastet, er sandsynligheden for, at terningen lander med et R opad 1/3. Hvor mange overflader er mærket R?” med svarmulighederne 5, 4, 3, 2 eller 1. I figur 2A er vist item karakteristik-kurven for denne opgave (item 106).

FIGUR 2A

Som det ses ligger de 5 punkter (som hver repræsenterer ca. en femtedel af de 207 elever sorteret efter location, det vil sige beregnet dygtighed) ikke nær den kurve, som de burde ligge på, hvis de skulle passe til Rasch-modellen. Punkternes placering fortæller, at dygtige elever ikke svarer mere rigtigt end mindre dygtige elever (målt ved Rasch-analysen) – nærmere tværtimod. Derfor er denne opgave ikke egnet til at indgå i testen. Hvad grunden er hertil, kan vi kun gætte på – et bud kunne være, at eleverne i slutningen af 6. klasse ikke har lært den form for sandsynlighedsregning, hvorfor det at svare korrekt beror på andre forhold end matematisk dygtighed.

Som det fremgår af tabel 1, er der udvalgt 94 items, hvor de 93 ligger inden for grænserne for fit-residualet. Det sidste item, som vi har valgt at inkludere i testen (item 60), overdiskriminer, hvilket er vist i figur 2B. Dette item har samtidigt den stærkeste korrelation (r = 0,56) med DNT af alle items, hvilket har medvirket til beslutningen om at lade det indgå.

FIGUR 2B (item 60)

Der er dermed udvalgt 94 items, som anvendes i den præ- og posttest, der skal anvendes til at følge elevernes progression i skoleåret 2014/15 på de fem skoler (nøgletal fremgår af tabel 2).

**Tabel 2**

| Korrelationer & antal items (N=175) | | DNT-mat6 | | | DNT i alt (beregnet) | Brutto items | Netto |
|---|---|---|---|---|---|---|---|
| | | T&A | Geo | MiA | | | |
| | | Korrelation | | | | Antal items | |
| NM-testen | T&A | 0,55 | 0,45 | 0,49 | 0,60 | 38 | 32 |
| | Geo | 0,55 | 0,57 | 0,50 | 0,65 | 34 | 29 |
| | MiA | 0,61 | 0,57 | 0,59 | 0,71 | 40 | 33 |
| | Samlet | 0,65 | 0,61 | 0,60 | 0,74 | 112 | 94 |

Det, man kunne have forventet, var, at diagonalen i tabellen havde vist de højeste korrelationer, men dette viser sig ikke at holde stik. Tværtimod er korrelationen mellem det overordnede resultat og de enkelte profilområde-resultater (på begge leder) i alle tilfælde højere end det, som fremkommer på diagonalen.

Årsagen er sandsynligvis, at definitionen af profilområder og klassifikation af items, som sker i hhv. DNT og i NM-testen (på grundlag af NAEP's klassifikation), er forskellig. Det betyder, at man kun med begrænset præcision (0,55 < r < 0,59) kan forudsige resultaterne i de enkelte profilområder i NM-testen ud fra oplysninger om resultater i DNT; mens man med noget større sikkerhed (r = 0,74) kan forudsige det samlede resultat ud fra den beregnede samlede DNT-score.

***Validitet***

Det er af særlig interesse, at vurdere om NM-testen og DNT måler samme underliggende egenskab. Da vi ikke har adgang til de enkelte item-responser fra DNT (altså hvilke opgaver den enkelte elev har besvaret, og om der er svaret rigtigt), men alene deres estimerede resultat af testen (locations), er vi henvist til at anvende redskaber fra den klassiske testteori.

Konkret kan vi gennemføre en ækvivalering[7], som vil give en indikation af, om det er den samme form for matematik, som måles. I så fald skal observationerne fordele sig jævnt omkring en ret linje med hældningen 45 grader. Af figur 3 fremgår sammenhængen mellem de to tests resultater. Sammenhængen er tydelig (r = 0,74), og der er indtegnet en OLS-regressionslinje.

FIGUR 3

Der er endvidere plottet en ækvivaleringskurve ind (de røde prikker), og som det fremgår, ligger den rimeligt pænt om en ret linje (den røde linje er regressionslinjen). Der hvor, der forekommer den mest systematisk afvigelse fra tendensen, er for de lavt præsterende elever. Det kunne tyde på, at fordelen ved det adaptive princip i DNT slår igennem, således at DNT især er bedre til at

bestemme de lavt præsterende elevers faglige niveau. Dette virker også sandsynligt, når opgaverne (som det fremgik af figur 1) gennemgående er ganske svære i forhold til elevernes dygtighed.

Samlet kan vi konkludere, at det ser ud til at være det samme konstrukt, som de to test måler, men da vi er henvist til at analysere på testscoreniveau (og ikke item-responser), kan vi ikke afprøve det yderligere.

***Reliabilitet***

Det næste vi skal se på, er den statistiske usikkerhed i forbindelse med målingen. I den klassiske testteori (CCT) opgøres sammenhæng mellem målefejlen (SEM - Standard Error of Measurement) og pålideligheden (reliabilitet) to modsat rettede størrelser.

Af Lord (1952) fremgår det, at:

- SEM = SD * kvadratrod (1 – reliabilitet)

Heraf følger, at

- reliabilitet = 1 – (SEM/ SD)^2

⇨ reliabilitet = 1 – SEM^2, når scoren er standardiseret (altså hvor variansen er normeret til 1)

En elevs testresultats statistiske pålidelighed (under de forudsætninger at testen måler det, den skal, og eleven gør sit bedste) afhænger især af to forhold: hvor mange items eleven har besvaret, og hvordan disse items sværhed passer til elevens dygtighed (”tagetting”). Jo flere målrettede items des større pålidelighed har målingen. Det fremgår ovenfor, at den statistiske pålidelighed kan opgøres enten som reliabilitet eller SEM (Standard Error of Measurement), som er udtryk for samme egenskab, men med modsat fortegn. Eftersom kun totalscoren i DNT er tilgængelig (vi har ingen oplysninger, om hvilke og hvor mange items den enkelte elevs DNT består af), kan vi ikke opgøre reliabiliteten i DNT.

Ved afvikling af DNT stilles en overvågningsmulighed, et monitoreringsskærmbillede, til rådighed for den lærer, som overværer testene. Herpå kan læreren se, hvor mange items hver elev har besvaret, samt om testen lever op til normale minimumskrav til testresultatets statistiske pålidelighed (vises med en farvekode der skifter fra rød, gul til grøn; hvor grøn indikerer, at testen overholder mindstekravet til reliabilitet). Tidligere har det været udmeldt[8], at grænsen for, at farvemarkeringen blev grøn, var en SEM på 0,3[9] svarende til en reliabilitet på godt 0,9. Dette er efterfølgende blevet korrigeret af Undervisningsministeriet, som oplyser, at grænsen for SEM ligger væsentlig højere (SEM=0,55)[10]. Dette modsvarer i dette sample en reliabilitet på 1-(SEM/SD)^2 = 0,71[11]. Vi kan se, at ca. 90 pct. af testresultaterne i vores sample har opnået farvekoden grøn, men der var altså 10 pct., som ikke levede op til reliabilitetsmålet.

For Rasch-modeller opereres der med en samlet reliabilitetsindikator PSI (Person Separation Index – leveres af programmet RUMM2030), og dette indeks skal helst ikke under 0,7. Hvis man skal have et estimat på personniveau, som tager højde for antal items og deres targetting ift. eleverne,

kan man anvende en reliabilitetsberegning fra den klassiske testteori, som er en fornuftig tilnærmelse (Thissen, 2000), selvom den ikke er teoretisk helt korrekt.

***Præ- og posttest***

De udvalgte 94 items er blevet fordelt på to test, hvor der er tilstræbt den samme fordeling af sværhed (ud fra Rasch-analysen) og faglige delområder (profilområder). Der arbejdes med et lille overlap (seks items bruges i begge test), hvorved både præ- og posttest bliver på 50 items og har PSI på henholdsvis 0,836 og 0,818.

Med udgangspunkt i afprøvningsdata er der undersøgt i, hvilket omfang de to nye test måler den samme egenskab. Dette er gjort på grundlag af responserne fra afprøvningen på 6. klasses-eleverne. Det er beregnet, hvordan de to test hver især ville blive scoret af eleverne. Opgjort som præ- og posttest er de to resultater plottet mod hinanden i figur 4.

Figur 4

Der er en markant større korrelation mellem præ- og posttestresultatet end mellem DNT og NM-testen (jf. figur 3). Korrelationen mellem præ- og posttest er dog lidt inflated på grund de 6 gengangeritems. Trækkes disse ud, bliver korrelationen reduceret en smule (r = 0,878) (jf. figur 3).

I figur 5 er det illustreret dels, hvordan resultaternes reliabilitet i NM-præ- og NM-posttest hænger sammen med elevernes præstation og antallet af besvarede items, og hvordan reliabiliteten af resultaterne fra DNT ligger i forhold til NM-testene. NM-prætesten ligger generelt højere end NM-posttesten, og begge test er targetted (dvs. måler mest præcist for) elever med en dygtighed (Rasch-score) på lige over 0. Kun for enkelte elever, som har besvaret væsentligt færre end alle items (punkter under "paddehatten"), kommer reliabiliteten under 0,70. Til sammenligning er indikeret et mål for reliabiliteten for DNT i nærværende sample af elever: 19 elever ud af de 175 (godt 10 pct.) opnåede ikke, at farvemarkøren for reliabilitet på monitoreringsskærmen gik fra gul til grøn. Når reliabiliteten er vist med en vandret linje, så er det for at vise, at DNT, pga. den adaptive udvælgelses af items principielt[12] fungerer, lige så godt for svage og stærke elever som for middelelever.

Figur 5

Det er dog vigtigt at være opmærksom på, at reliabiliteten øges ved at gennemføre flere items. Pædagogisk forskning (fx af Kristine Kousholt, AU/UIP)[13] har afdækket, at nogle skoler har den forestilling, at man bør stoppe elevens test i det øjeblik, at markeringen på monitorskærmen bliver grøn (at resultatet alternativt bliver ringere). Dette sker typisk efter 20-30 minutter og resulterer i, at elever, lærere og ledere på disse skoler får meget mindre sikre resultater at arbejde med, end hvad testsystemet normalvis vil kunne give, såfremt tiden blev udnyttet. Til brug for samarbejdsskoler er der udviklet en vejledning i, hvordan skolerne får retvisende testresultater[14].

***Om NM-test i matematik og egnethed til måling af progression på personniveau***

Det må på det foreliggende grundlag antages, at de konstruerede NM-præ- og posttest vil kunne belyse progressionen mindst lige så præcist som DNT. Baggrunden er, at disse test vil have 50 items placeret på én samlet matematikskala, som har en rimelig tilpasning til en IRT-model (enkelte items diskriminerer for kraftigt til at kunne passe på en Rasch-model). Til sammenligning

er DNT designet til at skulle måle tre separate skalaer, hvorfor der typisk er 15-25 items, som kan bidrage til beregningen af elevdygtigheden på hver skala. Det større antal items i NM-testene vil for langt de fleste elevers vedkommende mere end op veje målefordelen, som adaptiviteten i DNT medfører. Præcisionen i målingen af progression på personniveau afhænger af sikkerheden i de to målinger, som progressionen beregnes på grundlag af, hvor bidrag kommer fra henholdsvis:

1. Test-teknisk forhold
   a) antal items (jo flere besvarede opgaver des større præcision)
   b) items fit til skalaen (jo bedre fit til skalaen des større præcision)
   c) targetting (jo bedre items passer til elevens dygtighed des større præcision)
2. Personfaktorer
   a) kendskab til opgaveformater, vilkårene for testen mv. (jo bedre instruktion des større præcision)
   b) aktuel respons stil (den aktuelle dagsform, koncentration, omhu, jo mere eleven besvarer konsistent med sin latente dygtighed des større præcision (Buckley, 2009)
3. Den forventede faglige tilvækst over perioden
   a) Den normale faglige tilvækst pr. år (jo større progressionen er pr. år, des mere præcist kan den måles)
   b) måleperiodens længde (jo længere periode der måles over des større præcision). I denne undersøgelse er der tale om 11 måneder, da præstesten er gennemført i august 2014 og posttesten gennemføres i juni 2015.

Flere af disse forhold er velbelyste i denne artikel (1a, 1b og 1c samt 3b), andet vil der blive arbejdet med/blive afdækket i projektforløbet med skolerne i Aalborg kommune (2a og 3a), mens der vil være forhold, hvor vi er henvist til at skønne (2b).

Når der skal måles progression ved at gennemføre to test og se på differencen, så er der fejlkilder som reduceres (fx vil virkningen af forskelle i personlige forhold – herunder forskelle i personlig baggrund og respons stil – i betydeligt omfang neutraliseres), men samtidigt så forøges den maksimale målefejl. Derfor er det af væsentlig betydning at sikre en reliabilitet på et acceptabelt niveau. Samlet set vurderes der at være rimelige muligheder for statistisk signifikant at vurdere progression på personniveau på et overordnet niveau (fortegn, ingen, lille, stor), uanset at måleperioden er kort – i international forskningslitteratur anbefales en længere periode, fx 2-3 år (Haertel, 2013).

Når der aggregeres på klasse- eller skoleniveau (fx med henblik på at vurdere lærer- og skole-effekter), øges den statistiske sikkerhed samtidigt med, at der tilføres en betydelig kompleksitet (Darling-Hammond, 2012; Haertel, 2013).

## Diskussion: Virker reformen og hvordan kan man måle det?

Hovedformålet med reformen er at styrke elevernes læring. Det har været diskuteret, om der overhovedet kan forventes en sådan effekt, og hvor lang tid der skal gå, før man vil kunne se resultaterne i konkrete målinger. Der er argumenter i hver retning – hvad er der evidens for?

Evidens er blevet et begreb, som anvendes med mange betydninger. Det spænder fra, at man kræver, at der er bevidst kausalitet, til bare det er trykt i et peer-review'ed tidsskrift. Andre (fx

Hattie 2005) stiller krav om, at der skal kunne vises signifikant korrelation. I den oprindelige og mest restriktive form (”rigtig evidens” jf. Wainer, 2011) skal der tre ting til for, at man kan tale om evidens: 1) En god og velbegrundet hypotese (plausibel og til at afprøve), 2) en måling der viser signifikant og markant sammenhæng (korrelation) og endeligt, 3) at der kan vises en direkte årsagssammenhæng, hvilket indebærer, at konkurrerende forklaringer kan afvises (kausalitet). Disse skrappe krav er det så vanskeligt at leve op til inden for samfundsvidenskaberne, så ”rigtig evidens” meget sjældent forekommer inden for dette felt.

Det er især kausalitetskriteriet, der er vanskeligt at opfylde. Visse forskere er opmærksomme herpå (fx Hattie, 2012) og hævder ikke at finde kausalitet, men at arbejde med begreberne korrelation og sandsynlighed. I konceptet ”visible learning” giver Hattie fx udtryk for sine hypoteser, gør rede for at de er plausible og erklærer, at han ville holde fast ved dem indtil, at de bliver tilbagevist eller han bliver overbevist om stærkere forklaringer på det, som han kan observere (Hattie, 2012). Men der er i forskningsverdenen ikke enighed, om at alt, hvad Hattie giver udtryk for, er korrekt (jf. fx Snooks, 2009)

***Medfører reformen øget progression: En hypotese***

Effektforskning inden for uddannelse er et emne, man har beskæftiget sig meget med i de engelsktalende lande. Et af de mest kendte eksperimenter med effektmåling er ”The Perry Pre-school program” (Berrueta-Clement, 1984). Dette program bestod af en toårig indsats og stort set hele effekten kom i det første år; derefter forsvandt effekten gradvist (negativ progression) over en årrække (mange år senere vist der sig uventede resultater, men det er en anden sag).

Selvom effektvurdering i uddannelsesfeltet er relativt svagt belyst i norden, er der en række danske erfaringer at trække på:

- I december 2012 blev det i en artikel i Folkeskolen vurderet, at flere undervisningstimer ikke giver bedre elevresultater – der henvises til TIMSS- og PIRLS-undersøgelserne[15]. Heraf fremgår, at der ikke er nogen statistisk sammenhæng mellem det gennemsnitlige antal af undervisningstimer og landenes gennemsnitsresultater i internationale målinger. Det siger dog ikke nødvendigvis noget om effekten af en forøgelse af antal timer, som der er tale om med reformen.

- I et studie fra Aarhus Universitet fra 2014 fremgår det, at fire ekstra dansktimer om ugen i 16 uger giver en fremgang i læsning (dansk), der svarer til et halvt års normal undervisning.[16]

- Det mest omfattende forsøg på dansk grund er ”Tid til Dansk” fra 1990. Her sammenlignede man testresultater blandt danske elever, der havde fået en ekstra time i dansk om ugen i to år med kontrolelever, som ikke havde fået den ekstra time. På kort sigt var der en målbar, men begrænset effekt. En eftermåling viste, at kontroleleverne kort efter forsøgets ophør indhentede det læseforspring, som forsøgseleverne havde fået (Kreiner, 1992).

Erfaringer peger altså i retning af, at en forøgelse af timetallet kan give en effekt på kort sigt - men en mindst lige så vigtig faktor er, hvad der så gøres i de ekstra timer: mere af det samme eller noget andet?

Undersøgelser (Hattie, 2012) peger på, at det ikke er de ekstra undervisningstimer i sig selv, der giver resultater, men at man kan opnå betydelige effekter ved at benytte timerne rigtig. Hattie fremhæver, at elevers arbejde med opgaver/lektier på skolen med adgang til sparring og vejledning fra lærerside er en meget effektiv metode, hvor imod hjemmearbejde har ringe effekt.

Dette lægger sig ganske tæt op ad overvejelserne bag lektiecafeer; WEB'ben, som det kaldes på Sofiendalsskolen. Her er det intentionen, at eleverne skal have sparring og støtte i deres lektielæsning. Skolelederen refererer til diskussionen, om hvorvidt reformen skal være implementereret i flere år før, at den slår igennem på det faglige niveau[17]. Her giver lederen udtryk for, at han forventer, at man vil kunne se resultaterne hos eleverne i løbet af det første år[18]. Det, at lederen og lærerne tror på og arbejder for, at de nye måder at organisere sig på nytter, kan i sig selv have en betydelig effekt[19].

Når man taler om effekten af reformen, kan der være flere opfattelser af, hvad man egentlig taler om. Ofte betyder det implementering af de nye regler, arbejds- og tænkemåder i skolens liv (skolens interne organisering, lærernes måde at arbejde på, ledelsens rolle på skolen, arbejde med fokus på elevernes læring i stedet for egen undervisning m.m.). Det er sandsynligt, at der vil gå lang tid med at indarbejde nye vaner og arbejdsmåder, da fokus på undervisning (som paradigme, modsat læring, jf. Barr & Tagg, 1995) er et af de dybtliggende og markante træk i nordisk skolekultur (Chatterji, 2013).

I denne sammenhæng fortolker vi effekter af reformen synonymt med, om eleverne rent faktisk lærer mere, hvilket måles med NM-testen - og der er erfaringer for (jf. ovenfor), at der kan ske hurtige registrerbare ændringer i elevernes faglige niveau, når indsatsen forøges.

Samlet set kan ovenstående danne grundlag for at afprøve en hypotese om, *at den måde, som man har implementeret reformen på i undersøgelsens skoler, vil kunne give en signifikant effekt på elevernes progression allerede i det første skoleår 2014/15.*

***Hvordan kan hypotesen afprøves?***

Der er flere udfordringer heri, og de kan beskrives i tre kategorier. A) aggregeringsproblemet, B) demonstration af korrelation og C) eftervisning af kausalitet.

Ad A) Når der arbejdes med aggregering af elevers fremskridt sker det oftest for at kunne forbinde variationer i elevgrupper med skolens, lærerens eller kommunens indsats. Forskning i kommunens og skolens betydning (når der renses for lærereffekten) finder generelt ingen eller meget begrænsede effekter, og konklusionerne er rimeligt samstemmende (Ralph, Keller & Crouse, 1994; Hattie, 2005). Forskning peger på, at lærereffekten er den dominerende faktor i skolen, men at selv denne effekt er begrænset. At afgrænse, hvor stor del af elevernes udbytte, der kan henføres til læreren, er en meget omfattende diskussion, som fordrer avancerede statistiske modeller og stort datamateriale, hvis det skal gøres ordentligt. Dette udgør en selvstændig gren af engelsksproget uddannelsesforskning, som regel benævnes Value Added Modelling (VAM) eller Growth. VAM har fået en negativ klang i USA på grund af den måde, som model-data har været anvendt på.

Ad B) Vi vil kun overfladisk belyse forskelle i skolernes gennemsnit og spredning, og skolerne vil ikke kunne identificeres – i hovedanvendelsen kommer vi uden om de store teoretiske problemer i

aggregeringsøvelsen. Vi skal fokusere ind på progressionen umiddelbart før (i skoleåret 2013/14) og efter (2014/15) implementering af reformen. Vi vil måle progressionen fra starten til slutningen af skoleåret 2014/15 med NM-test, men vi har ikke fuldstændigt sammenlignelige oplysninger fra 2013/14. Der foreligger imidlertid oplysninger om resultaterne fra DNT for samme elever i perioden før reformen. Ved hjælp af den metode, som benyttes i Beregneren[20], kan resultaterne fra DNT omregnes til at vise elevernes progression på samme skala som NM-testene. Data, som kan inddrages, er derfor:

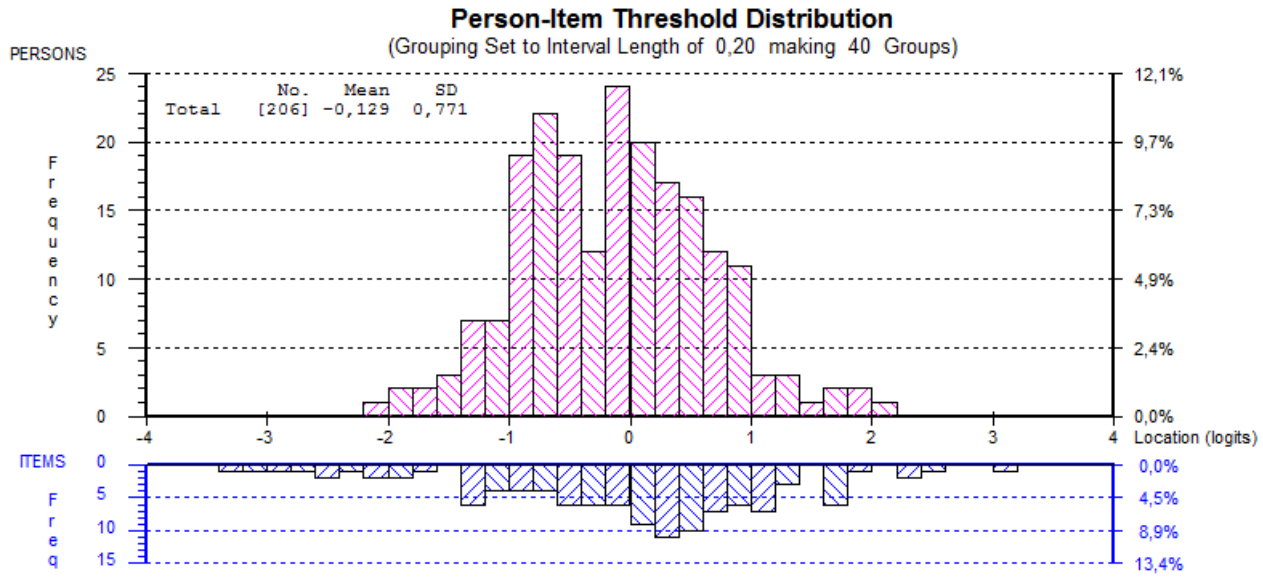
- Læsning (dansk): Oplysning om de nuværende 8. klasses-elevers (dvs. progression for 8. klasses-elevers i skoleåret 2014/15 fra 4. til 6. klasse og fra 6. til 8. klasse, men hvor sidste periode dækker både tiden før og efter reformen).
- Matematik: Oplysninger om de nuværende 8. klasses-elevers progression fra 3. til 6. klasse.
- Læsning (dansk): Oplysninger om de nuværende 9.klasses-elevers progression fra 6. til 8. klasse målt ved obligatoriske test i 2011/12 og 2013/14).
- Eventuel anvendelse af frivillige test kan endvidere undersøges.

Grundprincippet er, at progression måles som hældningen på en udviklings- eller læringskurve. Kunsten er at etablere så vidt muligt sammenlignelige udviklingskurver. Vi har den udfordring, at NM-testene ikke måler præcis på samme måde som DNT. Vi ved også, at den normale progression ændres hen over klassetrin, så vi kan ikke gå ud fra, at progressionen fra 5. til 6. klasse er halvdelen af progressionen fra 4.-6. klasse i et stort sample. Derimod kan vi måske ved at identificere mønstre i data komme med et kvalificeret skøn.

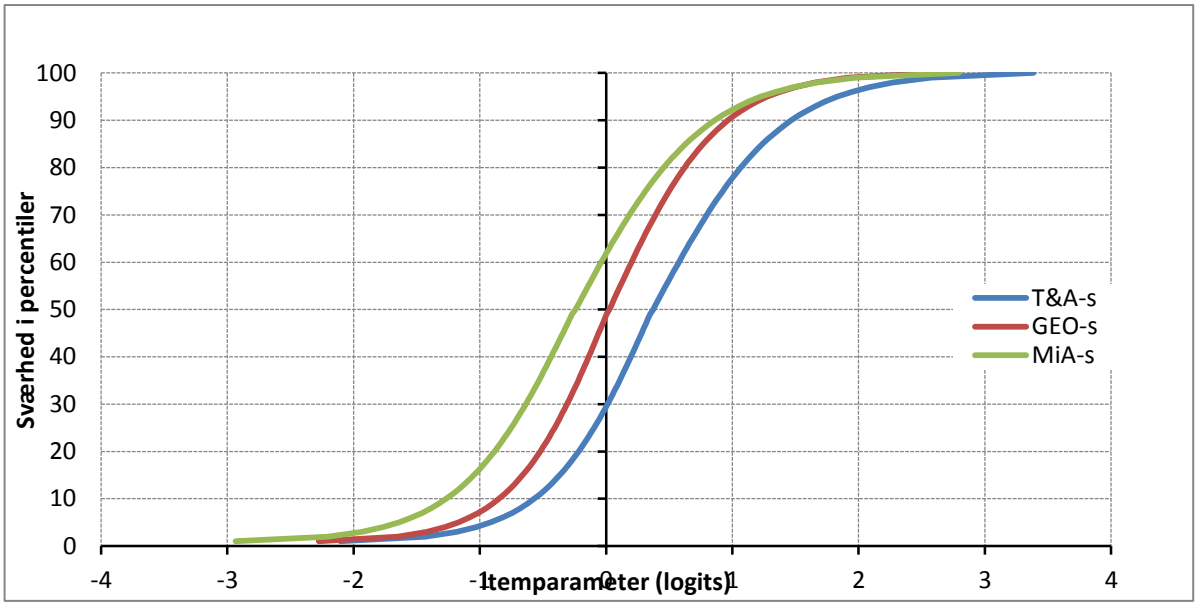
Der skal arbejdes nøjere med opstilling af statistiske modeller for sammenligning, og det skal afdækkes konkret, hvilke data der er til rådighed; før der kan siges noget mere præcist om dette forsøg på at måle effekten af reformen samt med hvilken statistisk sikkerhed, man kan sige noget om samvariation. Først herefter skal det overvejes, hvad der er virkninger af reformen og i hvilket omfang, det er muligt at isolere effekter fra andre faktorer (særligt arbejdstidsaftalen). Dette arbejde vil først kunne færdiggøres efter indhentning af data fra posttesten i juni 2015.

**Tabeller og figurer**

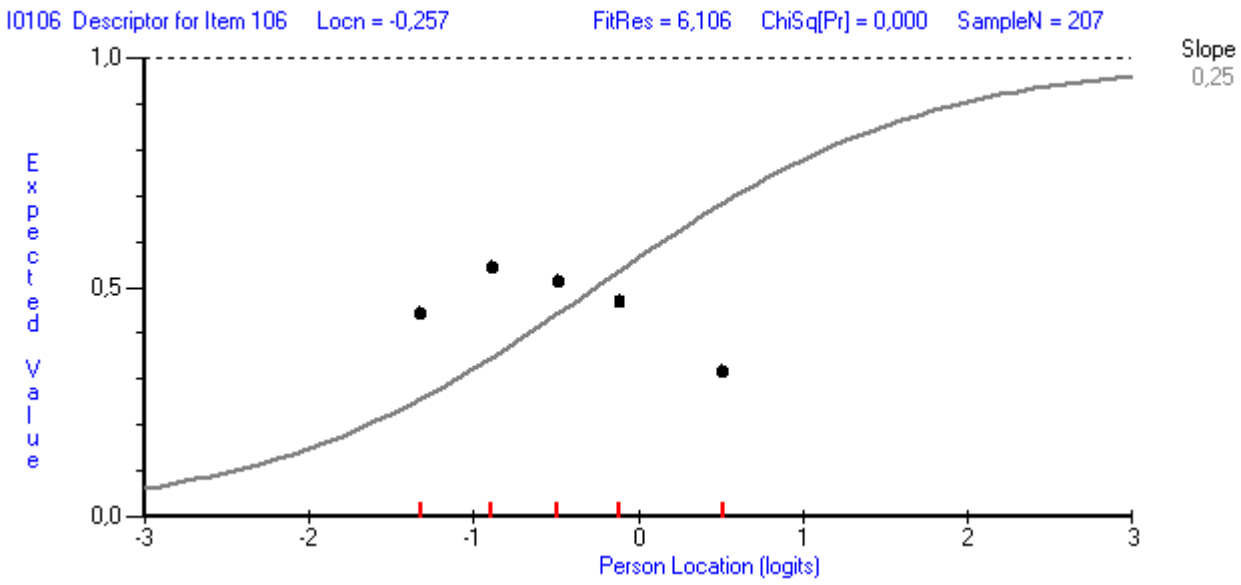
**Figur 1**



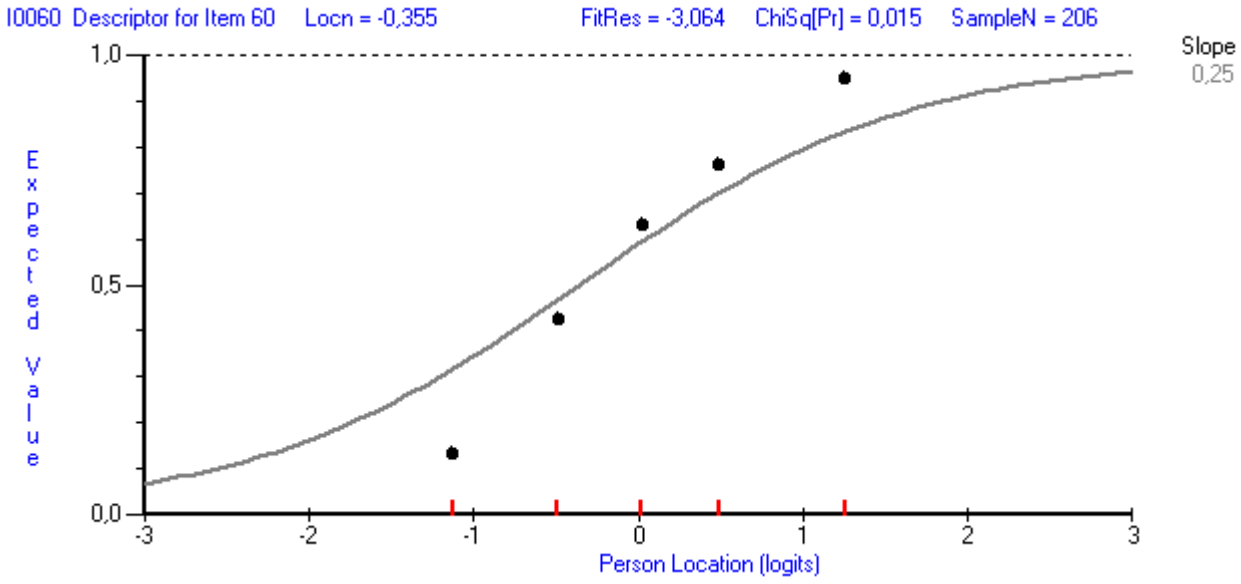
**Figur 2**


100
90
80
70
60
50
40
30
20
10
0
Sværhed i percentiler
-4
-3
-2
-1
0
1
2
3
4
Itemparameter (logits)
T&A-s
GEO-s
MiA-s

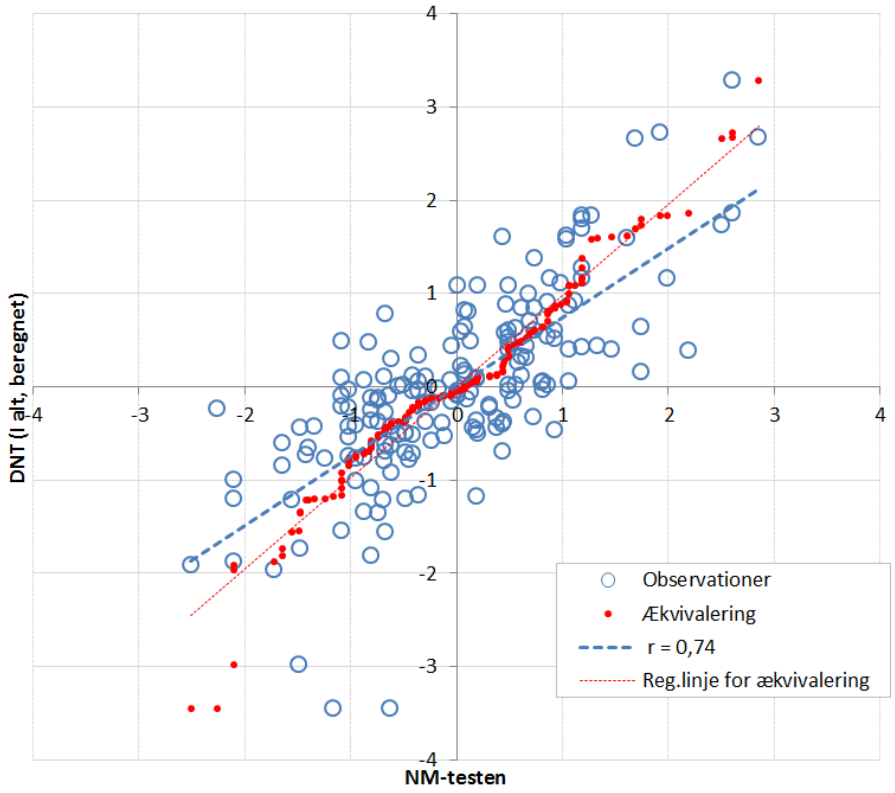
**Figur 2a**



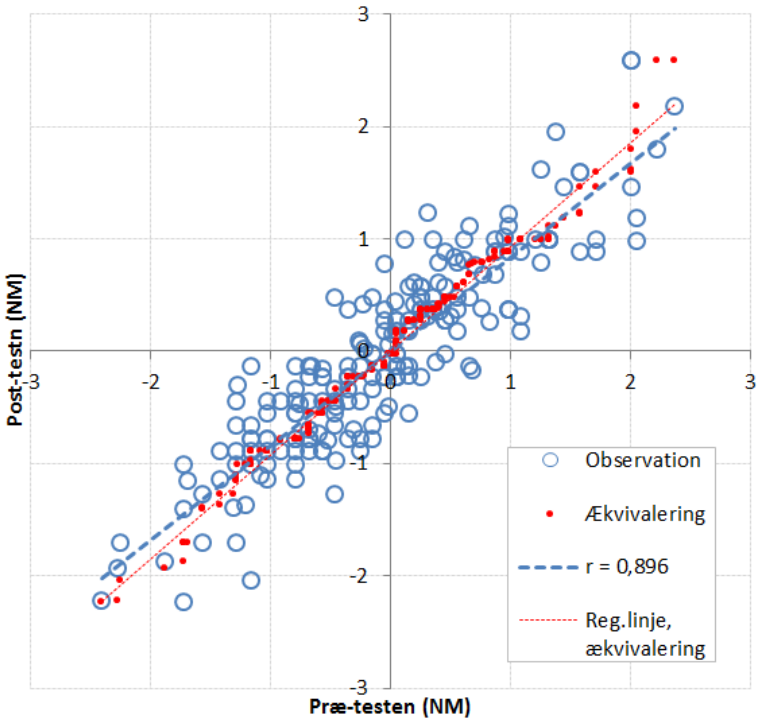

**Figur 2b**



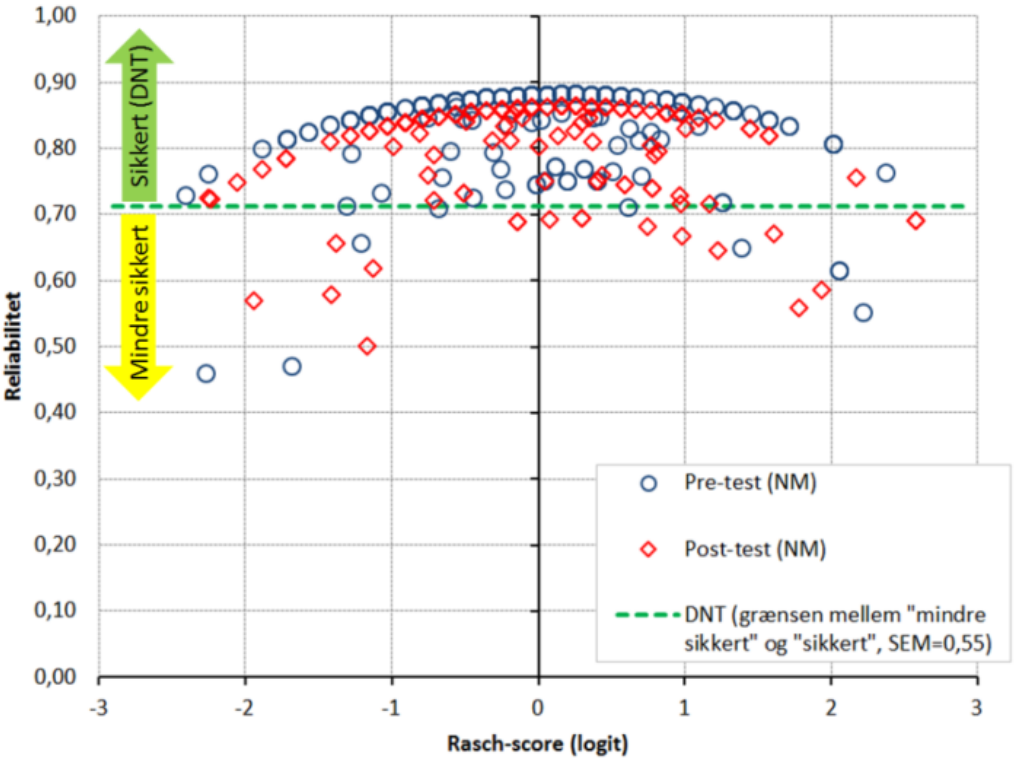
**Figur 3**


DNT (I alt, beregnet)
NM-testen
Observationer
Ækvivalering
r = 0,74
Reg.linje for ækvivalering

**Figur 4**


Post-testn (NM)
Præ-testen (NM)
Observation
Ækvivalering
r = 0,896
Reg.linje, ækvivalering

**Figur 5**


1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
Reliabilitet
Sikkert (DNT)
Mindre sikkert
-3
-2
-1
0
1
2
3
Rasch-score (logit)
Pre-test (NM)
Post-test (NM)
DNT (grænsen mellem "mindre sikkert" og "sikkert", SEM=0,55)

**Referenceliste**

Barr, R. B., & Tagg, J. (1995). From teaching to learning—A new paradigm for undergraduate education. *Change: The magazine of higher learning, 27(6)*, 12-26.

Berrueta-Clement, J. R. (1984). Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19. Monographs of the High/Scope Educational Research Foundation, Number Eight. Monograph Series, High/Scope Foundation, 600 North River Street, Ypsilanti, MI 48197.

Bonde, M. T., Makransky, G., Wandall, J., Larsen, M. V., Morsing, M., Jarmer, H., & Sommer, M. O. (2014). Improving biotech education through gamified laboratory simulations. Nature biotechnology, 32(7), 694-697.

Buckley, J. (2009, June). Cross-national response styles in international educational assessments: Evidence from PISA 2006. In NCES conference on the Program for International Student Assessment: What we can learn from PISA.

Chatterji, M. (2013). Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity. Emerald.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. Phi Delta Kappan, 8-15.

Dudley, P., & Swaffield, S. (2008). Understanding and using assessment data. *Unlocking assessment: Understanding for reflection and application.* Sue Swaffield (Edit). S. 105-120.

Egelund, N (red.) (2013). "PISA 2012-Danske unge i en international sammenligning". KORA, Det Nationale Institut for Kommuners og Regioners Analyse og Forskning.

Haertel, E. H. (2013). Reliability and Validity of Inferences About Teachers Based on Student Test Scores. *The 14th William H. Angoff Memorial Lecture, Washington, D.C.*

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J. (2005). What is the nature of evidence that makes a difference to learning?. *2005-Using data to support learning*, 7. ACER

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and writing, 2(2), 127-160.

Houlberg, K., Larsen, B. Ø., & Rangvid, B. S. (2013*). Benchmarking-og effektivitetsanalyse på folkeskoleområdet*. KORA.

Kreiner, S. (1992) Tid til dansk på folkeskolens mellemtrin: Om forholdet mellem undervisningstid og læseudvikling. Kongressrapport fra Den 11. nordiske lesekongress, *Norsk Forening for Leseopplæring*, pp. 28-40.

Lord, F. (1952). *A Theory of Test Scores.* Psychometric Monograph No. Richmond, VA.

Pøhler, L. (2010). Er de nationale test egnet til dine elever?. I: *Specialpædagogik, 2010, årg. 30, nr. 3,* s.19-24.

Ralph, J., Keller, D., & Crouse, J. (1994). How effective are American schools?. *Phi Delta Kappa*, 144-150

Rangvid, B. S. (2008). *Skolegennemsnit af karakterer ved folkeskolens afgangsprøver*. KORA.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30(2),* 175-194.

Snook, I., O'Neill, J., Clark, J., O'Neill, A. M., & Openshaw, R. (2009). Invisible Learnings?: A Commentary on John Hattie's Book-'Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement'. New Zealand journal of educational studies, 44(1), 93.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2005). How accurate are Lexile text measures? *Journal of* Applied Measurement, 7(3), 307-322.

Stenner, A. J., & Burdick, D. S. (1997). The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group. Ed.Gov. (US)

Thissen, D. (2000). Reliability and Measurement Precision. In Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy R. J. (Edit) (2000) *Computerized adaptive testing: A primer*. New York, Routledge

Undervisningsministeriet (2014a): *Nationale mål og enklere regler*. Nedhentet d.06.10 kl. 14:10 fra http://www.uvm.dk/Den-nye-folkeskole/Skoleledelse-og-styring/Nationale-maal-og-enklere-regler d. 29.09 2014

Undervisningsministeriet (2014b): *Bekendtgørelse af lov om folkeskolen*. Nedhentet d. 06.10 kl. 14:21 fra https://www.retsinformation.dk/forms/r0710.aspx?id=163970

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies*. Princeton University Press.

Wandall, J. (2011). National Tests in Denmark – CAT as a Pedagogic Tool. Association of Test Publishers, 12(1)

Williams, J. L. (2010). *Reading comprehension, learning styles, and seventh grade students.* School of Education, Liberty University

Slutnoter

[1] NordicMetrics er en privat forsknings- og udviklingsvirksomhed, som er etableret medio 2014. Yderligere information se http://www.nordicmetrics.com/om-os/

[2] Erfaringerne med klasseresultater fra de nationale test (fx data, indsamlet i forbindelse med udviklingen af Beregneren til de nationale test i læsning) viste, at denne afstand er normal i klasserne stort set uanset, hvilken skole, man ser på. Beregningen af afstanden foretages ved at sammenligne forskellen i rasch-scoren, for eleven der ligger i top og i bund i klassen (med 20 elever pr. klasse svarer det til en forskel på 90 percentiler – fx fra 5 % til 95 % percentilen) i de nationale læsetests profilområde tekstforståelse og holde det op mod den forbedring af testresultatet, der normalt opnås i løbet af et års undervisning af klassetrinnets elever på/omkring 50 % percentilen. For tekstforståelse viser resultaterne (på grundlag af resultater fra den elevgruppe, der i 2010 blev benyttet ved standardiseringen af percentilskalaerne i de nationale test, ca. 14.000 elever), at forskellen mellem top og bund opgjort på denne måde svarer til mellem 8,5 og 10,5 års progression, afhængigt af klassetrin. Oversigt over den gennemsnitlige spredning og progression på decentiler kan findes her (figur 10 på side 10): http://www.mitbuf.dk/sites/default/files/Vejledning_til_Beregneren-Progression_i_DNT_V3-4_0.pdf

[3] Items er velvilligt stillet til rådighed af MetaMetrics. Itemtypen og principperne bag scoringsmodellen er beskrevet mere detaljeret i Stenner et.al. 2005.

4 National Assessment of Educational Progress er en intern amerikansk sample baseret undersøgelse af amerikanske skoleelevers kundskaber og færdigheder i en række fag, heriblandt matematik. NAEP var metodisk en forløber for PISA. Flertallet af items i NM-testen er udarbejdet med udgangspunkt i offentliggjorte items fra NAEP, retrieved may 2014 from http://nces.ed.gov/nationsreportcard/itmrlsx/detail.aspx?subject=mathematics

[5] DNT er adaptive, hvilket betyder, at sværhedsgraden tilpasses efter eleven (Pøhler, 2010).

6 Det samme som standard-score, dvs. med en middelværdi på 0 og en varians på 1. Variansen er lig med kvadratet på standardafvigelsen, der ofte betegnes SD (Standard Deviation).

7 Med single group equi rank design, en særlig version af det ækvivaleringsprincip, der omtales som et single group equi percentile design i https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

8 På grundlag af oplysninger fra COWI, som stod for drift og udvikling af DNT

9 Dette er gengivet i Wandall, 2011

10 Se http://www.folkeskolen.dk/539694/ups--de-nationale-test-maaler-ikke-saa-praecist-som-lovet

11 SD =1,027 i nærværende sample.

12 Her forudsættes, at der er tilstrækkeligt mange egnede items i opgavebanken.

13 Fx forskning præsenteret på konferencen 3. april 2014, "Perspektiver på folkeskolens testpraksis", jf. http://edu.au.dk/forskning/video/testpraksis-04-2014/

[14] Se vejledningen her: http://nordicmetrics.com.linux13.curanetserver.dk/wp-content/uploads/2014/05/Hvad_enhver_laerer_boer_vide_om_at_gennemfoere_DNT.pdf

[15] Jf. http://www.fsl.dk/aktuelt/nyheder-debat/2013/2/professor-skolereformen-er-rent-spin/

[16] Jf. Simon Calmar, AU http://www.folkeskolen.dk/543566/forsoeg-flere-dansktimer-giver-bedre-laeseresultater

[17] Se fx http://www.dr.dk/Nyheder/Indland/2014/08/10/231541.htm og http://www.folkeskolen.dk/547768/skolelederformand-reformen-virker-foerst-om-flere-aar

[18] Skolelederne i på Sofiendalskolen i DR dokumentarserien "Folkeskolen – Forfra", første afsnit vist 7.10 2014.

[19] Rosenthal-effekten, også kaldet den selvopfyldende profeti, at menneskers adfærd påvirkes af de forventninger, som omgivelserne giver udtryk for.

[20] Beregneren til læsning (dansk), som er udarbejdet af NordicMetrics i samarbejde Københavns Kommune (se http://www.nordicmetrics.com/beregneren/ og http://www.mitbuf.dk/beregneren,

http://www.folkeskolen.dk/~/4/1/kronik-nationale-test-2.pdf samt http://www.folkeskolen.dk/538868/nyt-redskab-koebenhavns-laerere-kan-foelge-elevernes-udvikling-i-de-nationale-test