Til
Undervisningsministeriet
(Kvalitets- og Tilsynsstyrelsen)

Dokumenttype
Bilag til Evaluering af de nationale test i folkeskolen

Dato
September 2013

BILAG 4: SELVSTÆNDIGT NOTAT OM EKSPERT-VURDERING 2

RAMBØLL

SELVSTÆNDIGT NOTAT OM EKSPERTVURDERING 2

Rambøll
Hannemanns Allé 53
DK-2300 København S
T +45 5161 1000
F +45 5161 1001
www.ramboll.dk

BILAG 4
SELVSTÆNDIGT NOTAT OM EKSPERTVURDERING 2

Peter Allerup

EVALUERING AF DE NATIONALE TEST
- EKSPERTVURDERING 2

Indledning og formål

Formålet med dette notat er at udføre en vurdering af fordele og ulemper ved det såkaldt *adaptive princip,* som er en del af de nationale adaptive it-baserede test (herefter benævnt NT). For at foretage en sådan vurdering gives indledningsvis en kort beskrivelse af, *hvad* det adaptive princip rent pædagogisk og teknisk går ud på, og hvordan det forstås i lyset af den anvendte psykometriske metode i NT[1]. Den samlede vurdering er baseret på desk research og litteraturstudier. Empiriske erfaringer fra analyser med data fra den danske implementering af NT indgår ikke i dette notat.

De emner, som i forbindelse med evalueringen tages op, kan indledningsvis kort præsenteres i overskriftform:

1. *Teknisk grundlag for det adaptive princip*
2. *Pædagogisk egnethed af det adaptive princip*
3. *Fordele og ulemper ved det adaptive princip*
4. *Elevers og læreres erfaringer og oplevelser med det adaptive princip*
5. *Hensigtsmæssigheden af de nationale tests formidling*
6. *Måling af faglig progression*
7. *Det adaptive princip som grundlag for ny viden*
8. *Tekniske og økonomisk egnethed af det adaptive princip.*

[1] Det adaptive princip fortolkes forskelligt, afhængig af *hvilken* psykometrisk metode (statistisk model) der anvendes

Teknisk grundlag for det adaptive princip

Det adaptive princip er en procedure til udvælgelse af opgaver, som fungerer sammen med en elevs løbende besvarelser af testopgaver i NT. Der er ved NT tale om *præstations*opgaver i forbindelse med en række af folkeskolens almindelige fag, som for de fleste opgavers vedkommende besvares under det såkaldte *Multiple Choice* (MC)-format, dvs. opgaver hvor eleven kan vælge mellem et antal svarmuligheder, mere eller mindre præcist angivet på forhånd.

Det adaptive princip er en procedure, som aktivt medvirker ved udvælgelsen af den næste opgave, som eleven skal præsenteres for. Dette valg tager individuelt hensyn til de besvarelser, som eleven allerede har afgivet, og drejer sig om at vælge en opgave med passende sværhedsgrad som den næste opgave, når elevens dygtighed, beregnet ud fra afgivne besvarelser, tages i betragtning. Rammerne for det adaptive princip har to hovedparametre: Item*sværhed* og elev*dygtighed,* der skal spille sammen i en løbende vurdering og opdatering af elevens dygtighed estimeret ud fra besvarelse af forrige spørgsmål *og* sværhedsgraden af det næste spørgsmål. Modellen søger altså at udvælge spørgsmål, der kan fastlægge elevens niveau hurtigt. En teoretisk implementering af det adaptive princip i en formel analytisk model (statistisk model) må derfor finde sted inden for den specielle klasse af statistiske modeller, som går under betegnelsen IRT (Item Response Theory). Denne klasse er, modsat de "klassiske" psykometriske modeller, karakteriseret ved netop at inkludere begge aspekter; *sværhed* og *dygtighed* i den formelle struktur. Blandt IRT-modellerne findes de såkaldte Rasch-Modeller, som NT har anvendt som grundlag. Samtlige teoretiske overvejelser og vurderinger, fx vedrørende antal stillede opgaver, præcision ved bestemmelsen af elevdygtigheder mv. i NT, er dermed henvist til de teoretiske rammer, der er udstukket af IRT.

Den adaptive procedure består af en række delprocedurer og et stopkriterium:

1 Eleven stilles indledningsvis over for fem testopgaver, der bruges til en første beregning af elevens dygtighed. De fem opgaver vælges således, at enhver elev bliver stillet over for mindst én opgave, der er for vanskelig og mindst én opgave, der er så let, at den kan løses. Der startes med et middelsvært item. Når værdien af elevdygtighed *er lig med* værdien af itemsværhed, har eleven i følge IRT sandsynligheden 50 pct. for at svare rigtigt på den stillede opgave.

2 I de efterfølgende trin vælges testopgaverne successivt med sværhedsgrader, der ligger tættest på den værdi af den beregnede elevdygtighed, som kan opnås ud fra en vurdering af samtlige foregående svar. Dette gentages, indtil den statistiske standard error of measurement (SEM, som bliver systematisk mindre, jo flere opgaver eleven stilles over for) på beregningen af elevdygtigheden er reduceret så meget, at eleverne bør kunne nå at få et resultat for alle tre profilområder på 45 minutter. *Dette anvendes som stopkriterium*.[2]

Den måde, det adaptive princip er implementeret i NT, repræsenterer én blandt flere mulige metoder. Det gælder den måde, proceduren startes på, dvs. valg af indledende opgaver, det gælder metoden til at vælge "næste" opgave og det gælder det stopkriterium, som anvendes i forbindelse med beslutninger om, at eleven har fuldført testen. Der kan ændres på alle tre forhold uafhængigt af hinanden, og konsekvensen vil være, at eleven totalt set udsættes for flere eller færre opgaver i forbindelse med afviklingen af testen. Det har derfor også direkte konsekvenser for den tid, der bruges til at teste den enkelte elev. Der er praktiske omstændigheder og betingelser omkring afviklingen af NT, som man bør være opmærksom på, fx at der er tale om, at en *hel* klasse går samlet til skolens edb-lokale efter passende bookning af tid. Det betyder, at nogle elever bliver hurtigere færdige end andre, at den samlede testtid ønskes holdt inden for bestemte rammer etc., alt sammen forhold, som behandles på en måde, som skolen har fastlagt. Fx kan det være uhensigtsmæssigt, at de elever, der bliver hurtigt færdige, omgående forlader testlokalet efter endt testning. Et ønske om at benytte mulighederne for at ændre på de tre nævnte forhold skal derfor ses i sammenhæng med behovet for ændringer i forhold til de praktiske rammer for testafvikling, som skolen af andre end testmæssige årsager lægger fast.

[2] Det kan ikke lade sig gøre at implementere et stopsystem, som "virker" over for alle. Deraf formuleringen: "at eleverne bør kunne nå at få et resultat".

Teknisk grundlag for formidling af testresultater

Efter beregningen af elevdygtighederne defineres nogle grænser i den samlede fordeling over målene for elevdygtigheder, således at grupperne omfatter hhv. 10 pct., 25 pct., 30 pct., 25 pct. og 10 pct. af eleverne, kendt fra sædvanlig karaktergivning under 7-trins-skalaen (svarer til intervallerne [1-10], [11-35],[36-65], [66-90],]91-100], der igen svarer til *udmeldte niveauer 1-5*). Grænserne i den samlede fordeling er fastsat af populationen fra de første obligatoriske test i foråret 2010. Alle efterfølgende elevdygtigheder bliver vurderet i forhold til disse grænser. Elevdygtighederne transformeres til en skala fra 1 til 100 (percentilskala). Ved tilbagemeldingen til læreren vil tallet fra percentilskalaen blive oplyst på hvert profilområde og for testen som helhed. Sammen med denne opgørelse vil *elevens niveauplacering 1-5 blive oplyst*. Desuden vil besvarelsen af hver enkelt testopgave blive formidlet.

Resultaterne, som er umiddelbare produkter af den adaptive procedure, formidler information om *normative* sammenligninger, dvs. tal, som ikke umiddelbart giver information om, eleven er dygtig eller ej. De tal, som er resultatet af den adaptive procedure, skal forbi en fortolkning af ikke-maskinel art for at opnå samme status vedrørende udsagn om elevpræstationen, som karakterne fx i 7-trins-skalaen står for. Det er vigtigt at markere, at intet i NT's implementerede, adaptive rutiner giver adgang til fortolkninger vedrørende kvaliteten af elevens præstation i retning af "godt" eller "skidt". Den medfølgende karakteristik af elevens præstation ved hjælp af *udmeldte niveauer* indeholder ikke mere information end værdierne fra percentilskalaen. Hvis man ikke er tilfreds med normativt definerede vurderinger (altså rangordninger) og i stedet ønsker at formidle *simple målorienterede* evalueringsresultater fra NT, fx et ønske om at måle elevernes præstationer i simple procent korrektløste opgaver (som det er tilfældet ved andre typer af ikke-adaptive test fra folkeskolen), møder man den barriere, at eleverne i løbet af NT *ikke* løser de samme opgaver på grund af det adaptive princip, som tilpasser forskellige opgavesekvenser til forskellige elever. Der findes i litteraturen forskellige henvisninger til forsøg på at kreere et informationsgrundlag for målorienterede formidlinger af testresultater fra adaptive testsystemer[3]. Der trækkes i sådanne forsøg på lang tids erfaring med brug af informationer fra databanken og registrering af de formative tiltag og interventioner, som formidlingen til lærerne har givet anledning til. Et ønske om at få adgang til evalueringsresultater, der i højere grad forholder sig til de læringsmål, der ligger bagved, dvs. i højere grad bliver kriterieorienteret end tilfældet er nu, indfries ikke med NT. En kriterieorienteret vurdering betyder, at elevens evalueringsresultat knyttes til de læringsmål, der ligger bag opgaven med en didaktisk reference.

Pædagogisk egnethed af det adaptive princip

En vigtig side af vurderingen af den pædagogiske egnethed drejer sig om, hvor hurtigt og præcist man opnår pålidelige skøn over elevens dygtighedsniveau. Egnetheden af det adaptive princip skal derfor ses i lys af den indflydelse, det adaptive princip har på hastigheden, hvormed man beregner elevens "sande" dygtighedsniveau, samt den præcision hvormed dette estimat af dygtigheden er bestemt.

Hastigheden, hvormed elevens niveau fastsættes, afhænger af forvaltningen af det oven for beskrevne stopkriterium og af elevens adfærd ved besvarelsen af opgaverne undervejs i NT. Anskuet gennem *teoretiske* IRT-briller (idet dette notat er en vurdering af fordele og ulemper ved det adaptive princip baseret på litteraturreview) må man skelne mellem to typer af elever: (1) Elever, der udfordres af "næste" opgave med et svar, som kan prædikeres ved hjælp af den bagomliggende statistiske IRT-model og (2) elever, som lader sig inspirere af alt muligt andet end selve opgaven, når de skal svare.

[3] *Item selection and ability estimation in adaptive testing* (Wim J. van der Linden and Peter J. PashIey). *Adaptive tests for measuring anxiety and depression* (Otto B. Walter). *Designing item pools for adaptive testing* (Bernard P. Veldkamp and Wim J. van der Linden).

Den første gruppes hastighed eller samlede testtid kan beregnes ud fra den statistiske IRT-model, og al litteratur viser[4], at det adaptive princip fører til testtider, som er lavere end den tid, der afvendes ved sædvanlig, Traditionel Lineær Testning[5] (TLT – papirbaserede test). Kun i det ekstreme tilfælde, hvor en elev på en sædvanlig lineær test afleverer "blankt" = fejl på *samtlige spørgsmål,* eller er så dygtig, at der umiddelbart svares rigtigt på samtlige stillede opgaver, kan der blive tale om, at det adaptive princip evt. forlænger testtiden. Men for disse to ekstreme grupper opnås til gengæld et numerisk estimat af elevens dygtighed via den adaptive procedure - som enten tilordner lettere og lettere eller sværere og sværere spørgsmål, indtil elevniveauet bestemmes – et estimat, som ikke er tilgængeligt ved sædvanlige lineære test.

Den anden gruppe elever udgør et større problem ved de adaptive test sammenlignet med sædvanlige lineære test, hvor opgavemængden er bestemt på forhånd. Denne gruppe af elever kan forhale testtiden principielt i det uendelige, fordi eleven er mere optaget af at "drille" systemet end at løse opgaverne. Den samlede testtid for disse elever kan derfor let overstige den tid, der ellers anvendes ved lineære test.

Det adaptive princips *indflydelse på præcisionen* af bestemmelsen af elevdygtigheden er allerede beskrevet via det adaptive princips stopkriterium. Dette forudsætter, at den adaptive proces fortsættes, indtil den statistiske *standard error of measurement* (SEM) blive reduceret *så meget, at eleverne bør kunne nå at få et resultat for alle tre profilområder på 45 minutter*. En definition, som ikke er så simpel og klar, at den sætter absolutte grænser for SEM, men som i lys af IRT-modellen, vurderet ud fra rent teoretiske vinkler, *må antages* i praksis at føre til faste grænser, som er operationelle ved afviklingen af selve testen[6].

Den opnåede præcision ved testen af en elev indeholder imidlertid også et andet aspekt i form af de opgaver, som eleven konkret har været igennem. Umiddelbart skulle man tro, at elever via det adaptive princip bag udvælgelsen af opgaverne kan risikere "kun" at blive testet i et begrænset antal læringsmål (fordi udvælgelsen af næste opgave i høj grad er fokuseret på opgavens sværhedsgrad – og mindre på det specifikke faglige indhold bortset fra de faglige overskrifter, som de tre profilområder definerer). Modsat, skulle man derfor mene, kan man i de lineære test præcist bestemme, hvilke læringsmål der ligger bag ved testningen ved valget af opgaver og dermed komme "helt omkring" mht. at teste de samlede læringsmål. Dette ræsonnement er forkert, fordi den opbygning af databanken, som er gået forud, har været et filter for opgaver, hvor det netop er et kriterium (det såkaldte homogenitetskriterium) for indlemmelse i databanken, at en hvilken som helst opgave kan erstatte en hvilken som helst anden opgave – *alle* opgaver trækker på samme latente dimension. I denne beskrivelse er begrænsningen, at der tænkes på *"opgaver inden for samme profilområde"*. Der er fx inden for matematik tre latente områder: Tal og algebra, geometri og matematik i anvendelse. Samtlige opgaver inden for hvert af områderne adskiller sig *alene* ved deres sværhedsgrad, og altså ikke ved deres indhold (eller reference til læringsmål). Man kan konkludere, at foreningen af det adaptive princip med konstruktionsteknikken bag ved itembanken sikrer, at *faglig præcision i den beskrevne forstand* er den samme, som man finder ved tilsvarende lineære test.

Som omtalt efterspørges evalueringsresultater, som i højere grad end tilfældet er nu, eksplicit reflekterer elevens præstation i relation til de læringsmål, der ligger bag ved opgaven. Der ønskes med andre ord en type evaluering, som via en mål- eller kriterieorienteret evaluering meddeler, i hvor høj grad eleven har indfriet de krav, der er formuleret i læringsmålene (Fælles Mål). Ønsket svarer i høj grad til de oprindelige intentioner bag ved 7-trins-skalaen, hvor selve karakteren i princippet gives ved at tælle ned fra karakteren 12, skridt for skridt afhængig af hvilke og hvor mange fejl eleven har begået. Der vil med en stærkere fokus på kriteriebaserede evalueringer opstå samme problemer som ved karaktergivningen i starten efter 7-trins-skalaen, som består i,

[4] *Practical Implications of Item Response Theory and Computerized Adaptive Testing* http://www.jstor.org/stable/10.2307/3768065 *Principles of multidimensional adaptive testing* (Daniel O. Segall).
Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4

[5] En lineær tests er i denne forbindelse betegnelsen for en test, der er bygget op på forhånd af et fast sæt opgaver, som samtlige elever besvarer.

[6] Der indgår, som omtalt, ingen empiriske data i nærværende evaluering.

at man opgiver at kontrollere, hvor mange elever der tildeles de forskellige karakterer på skalaen. Med et fravalg af normrelaterede evalueringer til fordel for kriteriebaserede evalueringer bliver den simple forståelse besværliggjort af et "landsgennemsnit".

Fordele og ulemper ved det adaptive princip

For at kunne vurdere hvilke fordele og ulemper det adaptive princip (i forhold til lineære test) indebærer for testningen, er det nødvendigt at præcisere, at sådanne afvejninger afhænger af den psykometriske metode (model), som er anvendt ved den faktiske implementering. Altså i dette tilfælde den IRT-Rasch-Model, som har været anvendt ved konstruktionen af NT. Fx kan specifikke *elev*relaterede forhold diskuteres og analyseres under IRT, noget som ikke lader sig gøre i lys af de klassiske psykometriske modeller, hvor den individuelle elevs latente dygtighed *ikke* er en del af parameterstrukturen.

Fordele og ulemper af det adaptive princip bør vurderes i forhold til opgavebanker, med eller uden egenskaben af at være Rasch-homogene[7]. Under alle omstændigheder kræver implementering af et vilkårligt adaptivt testsystem tilstedeværelsen af en opgavebank. Det kan betragtes som en ulempe i forhold til sædvanlig lineær testning, at oparbejdelsen og vedligeholdelsen af en sådan opgavebank kalder på omfattende arbejds- og analyseresurser for at være et validt grundlag for en adaptiv procedure. Modsat traditionelle lineære papirbårne test (TLT) kræver udarbejdelsen af opgaver til opgavebanken bag NT flere arbejdsfaser:

I udviklingsfasen udvikler medlemmer af opgavekommissioner forslag til opgaver. Disse kvalitetssikres via eksterne kvalitetssikrere samt embedsmænd i ministeriet. Denne proces vil være fælles for NRT og TLT.

I pilotfasen afprøves opgaverne på en stikprøve af elever på relevante klassetrin, der observeres, mens opgavebesvarelsen finder sted. I nationale test sker dette blandt et begrænset antal elever stort nok til at sikre, at afprøvningen foregår på et validt niveau. Udvalgte elever interviewes omkring den tænkning, som ledsager løsningsprocesserne. Denne proces vil også være fælles for NRT og TLT.

I afprøvningsfasen, internationalt kaldet Field Trial (fx fra OECD's PISA samt IEA's TIMSS og PIRLS undersøgelser), afprøves betingelserne for den senere prøvning – både praktiske aspekter vedrørende prøveafholdelsen og teoretiske aspekter ved opgavernes egenskaber tilhører denne fase.

Opgaverne pakkes og udsendes under hensyn til nøje definerede kvalitetssikringsprocedurer, og afprøves på skolerne.

De løste opgaver returneres og udsættes herefter for Rasch analyse i Uni-C, hvor de psykometriske egenskaber nøje vurderes i forhold til kravene i RASCH modellen (se nedenfor).

Ved TLT blev fasen 'forcensur' efterhånden erstattet af en fase, hvor den egentlige prøveafholdelse blev lagt til grund for fx omsætningstabeller ved karaktergivning. Ved NT benyttes fasen til at indsamle viden om opgavernes psykometriske egenskaber. Denne fase fylder meget i NT og skal rent praktiske samtænkes med en praktisk løsning på det problem, at opgavebanken løbende justeres samtidig med, at der afholdes egentlige prøver (hvorunder opgaver i denne fase jo ikke tæller med).

I analysefasen, i de internationale undersøgelser kaldet Main Study fasen, benyttes de afprøvede rutiner som grundlag for selve prøveafholdelsen. Ved NT foregår dette ved hjælp af computere og det programmel, som er udarbejdet til NT, mens afviklingen af TLT kræver resurser til tryk, ud-

[7] Hvilket konkret betyder, at opgaverne har været underkastet psykometriske valideringer ud fra en IRT-model, fx en Rasch-model, og er blevet "godkendt" via de dertil knyttede statistiske test.

deling, indtastning og efterfølgende dataanalyse med henblik på konstruktion af omsætningstabeller som grundlag for den endelige karaktergivning.

Uni-C analyserer de indsamlede besvarelser med Rasch-analyse for at fastsætte opgavernes psykometriske egenskaber, og om opgaverne lever op til alle krav for at kunne indgå i opgavebanken.

Især dette er stærkt arbejdskrævende og fordrer statistisk faglig indsigt i analyseresultater ud fra kontrol af IRT-modellen. En del (måske op til 50 pct.) af opgaverne udelades, fordi de ikke besidder de ønskede psykometriske egenskaber. Nogle af disse kan revideres og afprøves efter samme proces, som ovenfor beskrevet. Forkastede opgaver eller opgaver, der ønskes ændret, selv meget små ændringer i overskrift og lignende, kan gennemgå et ændringsarbejde efter samme proces, som beskrevet i ovenstående skridt.

Selve arbejdet med udarbejdelsen af supplerende opgaveforslag kan varetages af opgavekommissioner inden for relevante fag/klassetrin. Det må betragtes som en ulempe, at sådanne personer skal være i stand til at forstå og aktivt handle på tilbagemeldinger fra den midterste fase beskrevet ovenfor, afprøvninger af nye items[8]. Der nævnes i litteraturen muligheden for at medtage nye opgaveforslag i NT blandt de eksisterende opgaver i opgavebanken. Altså at blande velafprøvede opgaver med nye opgaver. På denne måde opbygges en række svar på nye opgaver, som i forbindelse med svar på andre, gamle opgaver kan udgøre et grundlag for at vurdere de psykometriske egenskaber ved de nye opgaveforslag. Det sker på en sådan måde, at eleverne ikke "opdager" de nye opgaveforslag, idet de ikke medtages ved udregningen af elevernes dygtighed.

Aktiviteter og økonomi knyttet til opbygning og vedligeholdelse af opgavebanken som grundlag for det adaptive princip er beskrevet senere.

Den mest markante måde, det adaptive princip sørger for en forskel til traditionel lineær testning, ligger i, at *alle* elever vil opleve, at de har løst ca. 50 pct. af de stillede opgaver korrekt. Dette faktum er et produkt af det adaptive princip og det kan rent teknisk uden problemer styres til at ligge på et andet niveau end netop 50 pct., hvis det ønskes. Der synes at være enighed om, at elever har det bedst i testsituationer, når deres chance for at løse en (præstations)opgave ligger et sted mellem 20 pct. og 80 pct.. Uden for dette interval kreeres enten frustration eller kedsomhed hos eleverne. Det anvendte niveau på 50 pct. ligger, som det ses, lige i midten mellem 20 pct. og 80 pct., men ligger lidt under de ca. 60 pct., som normalt anvendes ved internationale testninger, PISA og IEA's test for eksempel. Der er i litteraturen[9] enighed om, at det især er de svage elever, som har gavn af det adaptive princip, forstået ud fra psykologiske/pædagogiske synsvinkler. Den omstændighed, at man løbende kan tilpasse opgavesværheden, afstedkommer en øget accept af testsituationen specielt for denne gruppe elever, sammenlignet med sædvanlige lineære test. De dygtige elever, som normalt kan løse de fleste af de opgaver, de stilles, kan opleve den effekt, at man tager modet fra dem mht. til selvforståelse af deres faktiske faglige niveau. Det er derfor vigtigt at tydeliggøre og forklare det adaptive princip over for eleverne, før de går i gang med testningen[10].

Som omtalt er en konsekvens af det adaptive princip, at eleven konstant præsenteres for opgaver, hvis sværhedsgrad matcher elevens dygtighedsniveau. Rent teknisk har dette den konsekvens, at den statistiske information målt via variansen bliver maksimal. Det adaptive princip betyder, at der opgave for opgave successivt opbygges viden om elevens dygtighedsniveau, således opbygget med maksimal information. Derfor er det adaptive princip alt andet lige de lineære test overlegen mht. at fremskaffe sikker viden om elevens dygtighedsniveau ved hjælp af færrest

[8] *Field Trial* gælder den fase af afprøvningen af nye opgaver, hvorunder et antal elever besvarer forslag til nye opgaver samtidig med besvarelsen af et antal "gamle" opgaver. Besvarelserne analyseres ved IRT statistiske modeller (Rasch-modeller), og en række *test statistics* formidler oplysninger om tilfredsstillende eller ikke-tilfredsstillende tilpasning til IRT-modellen. Disse oplysninger benyttes til at afgøre, om opgaveforslaget forkastes, revideres (i overensstemmelse med test statistics) eller accepteres.

[9] *Designing item pools for adaptive testing* (Bernard P. Veldkamp and Wim J. van der Linden).

[10] *Brugervejledning Testsystemet – De nationale test (UNI-C, januar 2013)*

mulige opgaver[11]. Det er en erfaring, at adaptiviteten medvirker til en ca. 50 pct. reduktion i mængden af opgaver, som er nødvendige at stille under disse betingelser sammenlignet med sædvanlige lineære test[12].

Det adaptive princip er uløseligt forbundet med eksistensen af en opgavebank (itembank) af en rimelig størrelse. Jo større banken er, jo mindre er sandsynligheden for, at samme opgave udtrækkes til to elever i samme klasse i samme testsession. Det var et vigtigt implementeringskrav for NT, at denne sandsynlighed skulle være lille. Sandsynligheden for, at samme opgave udtrækkes to gange, er kun delvist afhængig af antallet af opgaver pr. profilben (se appendiks), og er i højere grad afhængig af den metode, der ligger bag ved udvælgelsen af "næste opgave". Det skal i denne forbindelse fremhæves som en fordel, skabt af det adaptive princip, at man kan gentage adaptiv testning *uden* at lade eleverne (eller lærerne) få fornemmelse af genkendelighed af opgaverne. Det kan betragtes som en ulempe, hvis to elever, der har samme beregnede dygtighedsniveau, taler sammen om, at de undervejs har fået samme opgave, men at de har besvaret den forskelligt.

De økonomiske omstændigheder ved vedligeholdelse og opdateringer af opgavebanken behandles sidst i notatet. Sammenlignet med traditionelle lineære test er der bred enighed om, at IRT-definerede adaptive testsystemer kræver den største arbejdsindsats i forhold til TLT[13]. Det skal understreges, at der her tænkes på IRT-relaterede adaptive systemer, hvor kun opgaver, som alle har passeret forskellige statistiske tjek for itemhomogenitet (ud fra Rasch-modeller), indgår i opgavebanken (som den danske model). Der findes adskillige adaptive testsystemer, som opererer med opgavebanker, der *ikke* er bygget op af homogene items. Og derfor ikke er mere krævende mht. arbejdsresurser end sædvanlige lineære test. Fordelen ved at operere *med* IRT-relaterede opgavebanker over for itembanker *uden* homogenitet består i, at de sammenligninger mellem elever (og grupper af elever), der udføres med IRT-relaterede opgavebanker, er valide og kan gennemføres på én fælles (latent) skala, modsat sammenligninger, som er udført ud fra ikke-homogene opgavebanker. Når der er tale om faste præstationsskalaer, kan der derfor foretages meningsfulde sammenligninger fra år til år, som sædvanlige TLT-prøver ikke er i stand til. Ved TLT er sammenligningerne ikke bedre (valide), end de sammenligninger man kan foretage ud fra resultater fra to helt forskellige test, fx en traditionel (lineær) matematikprøve stillet til 9. klasse ét år i relation til den matematikprøve, som stilles det følgende år. Adaptiviteten sikrer hermed gennem kravet til opgavebanken under NT en stor psykometrisk fordel, sammenlignet med ikke-adaptive testsystemer, også i forhold til adaptive testsystemer bestående af ikke-homogene opgavebanker.

Under den praktiske afvikling af NT opbygges der opgave for opgave et stadigt klarere billede af elevens dygtighedsniveau. Det samme sker, hvis eleven starter en traditionel lineær prøve med opgave nr. 1 og derefter fortsætter opgave for opgave. I den traditionelle lineære version modsvares beregningen af elevens latente dygtighedsparameter af successivt beregnede procent korrekt, udregnet i forhold til det på tidspunktet passerede antal opgaver. Det er imidlertid den markante forskel mellem de praktiske prøveomstændigheder ved de to prøveformer, at ved traditionel lineær testning kan eleven "gå tilbage" og rette et afgivet svar. Det kan ske undervejs i prøvesituationen, eller når eleven sidder og har mere tid efter besvarelsen af sidste opgave, altså fx er blevet færdig meget før den planlagte testtid. I dette lys må det anses for en mangel ved NT, at eleven er uden mulighed for at rette tidligere afgivne svar. Stopkriteriet og forståelsen af den løbende allokering (adaptiviteten) af nye opgaver umuliggør dette. Det rejser på den anden side *ikke* nye principielle problemer ved selve beregningen af elevens dygtighedsniveau. Man kan sige, at besvarelse af opgaven, set i bakspejlet, kan betragtes som en version af en traditionel lineær test, stadig bestående af Rasch-homogene opgaver.

---

[11] Cees A. W. Glas and Hans J. Vos.*Adaptive mastering testing using a multidimensional IRT model /*Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4

[12] Wim J. van der Linden and Peter J. PashIey*Item selection and ability estimation in adaptive testing /*(Cees A.W. GIas).*Item parameter estimation and item fit analysis*

[13] Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4

Det adaptive princip fordrer som omtalt eksistensen af en itembank. Lægges der vægt på, at elever på samme dygtighedsniveau i en normalklasse *ikke* undervejs i deres testforløb udsættes for samme opgave, er det indlysende, at antallet af opgaver i midterområdet, der "passer" til de fleste elever (hvis dygtighedsmål fordeler sig med tyngden i midten), rent sværhedsmæssigt skal være betydeligt større end antallet af opgaver med enten lave eller meget høje sværhedsgrader.

Derudover er der hensyn at tage mht. størrelsen af itembanken, hvis dette krav om manglende genkendelighed af opgaver kædes sammen med ønsket om at øge itembanken med nye opgaver. I denne situation støtter det adaptive princip fornemmelsen af, at elever ikke kan udveksle erfaringer mht. indhold af opgaver (den manglende genkendelighed), og der skaffes et grundlag for at "pilotere nye opgaver ved at lade dem indgå mere eller mindre tilfældigt i rækken af godkendte opgaver – uden at tælle med ved bestemmelse af elevens dygtighedsniveau. Det ser ud til, at et minimum af ca. 1000 elevers svar på sådanne opgaver er nødvendigt som datagrundlag for en senere stillingtagen til, om den nye opgave kan indlemmes i itembanken eller ej. Indeholder datagrundlaget betydeligt færre elever, svækkes styrken af de statistiske test, der lægges til grund for analyser af tilpasning til Rasch-modellen (itemhomogenitet).

Mens fordelene ved det adaptive princip således kort kan opgøres som øget præcision under brug af færre opgaver, bedre testtrivsel over for eleverne og valide elevsammenligninger, er den markante ulempe, at det adaptive princip kalder på betydelige arbejdsresurser. Eller, som der skrives i en artikel[14]: *"Der er mange fordele ved at bruge adaptive test som i Danmark. Der opnås eksempelvis en god og effektiv testsituation, testen kan genbruges og testen giver sandsynligvis et mere retvisende billede af elevens kompetencer og kvaliteten i undervisningen. En ulempe ved de adaptive test er, at de er dyre at udvikle".*

Elevers og læreres erfaringer og oplevelser med det adaptive princip

Det ligger i den måde, det adaptive princip fungerer på, at når en elev præsenteres for "den næste opgave", vælges den tilfældigt blandt en række opgaver (inden for samme profilområde) med identiske sværhedsgrader. To elever, som har samme estimerede dygtighed, står derfor over for forskellige opgaver ved "næste opgave", afhængig af udfaldet af det tilfældige valg. Der er en risiko for, at denne side af det adaptive princip kan afstedkomme en følelse af forvirring hos eleverne, som evt. oplever gennemførelsen af NT som en "hoppen rundt" fra emne til emne – selv om opgaverne rent faktisk er valgt inden for samme profilområde. Det er normal praksis ved udformningen af traditionelle opgaver (TLT), at man undgår at "blande" indholdet i successivt stillede opgaver for meget.

Hermed kommer operationaliseringen af det adaptive princip til at modarbejde et grundlæggende princip, som ellers anvendes ved konstruktioner af test, hvor eleverne præsenteres successivt for en række opgaver: Der skal være en *indre sammenhæng*. Ved traditionelle lineære test (Folkeskolens afgangsprøver mv.) består én af dyderne for erfarne opgavekonstruktører bag prøverne netop i, at eleverne trods forskellighed i opgaverne hurtigt føler, at man befinder sig inden for et bestemt fagligt domæne. En eksplicit anvendelse af denne "sammenhængsteknik" ser man fx ved de internationale læseprøver (PISA), hvor en lang række opgaver henviser til samme grundlæggende tekst – som hermed er *det* sammenhængende hele, der åbenbares for eleven. Tilsvarende sammenhæng er fremhævet som noget positivt ved visse matematikopgaver, hvor især geometriske opgaver ofte anvender samme grafiske grundkonstruktion, der herigennem skaber den fornødne sammenhæng. De gængse psykologiske evalueringsinstrumenter (såkaldte rating scales, fx af depression og angst) kalder i ekstrem grad på, at respondenten oplever de stillede spørgsmål, hvad enten det foregår adaptivt eller ej, som dele af én kognitiv helhed.

Disse hovedsageligt negative erfaringer med det adaptive princip modsiges af andre, mere tekniske argumenter og erfaringer. Disse argumenter er lejret i en statistisk grundforudsætning ved-

[14] http://www.skoleraadet.dk/sitecore/content/Skoleraadet/Aktiviteter/Seminarer/Seminarer%202007/~/media/Skoleraadet/ Aktiviteter/Seminarer/Seminarer%202007/Foraar/julius%20bj%C3%B8rnsson%20referat.ashx

rørende besvarelsen af opgaverne i NT: *Lokal uafhængighed*[15]. Groft sagt må et svar på et spørgsmål ikke være påvirket af, hvorledes man har svaret på et andet (foregående) spørgsmål. Det var et resultat af det første review af NT, at kravet om lokal uafhængighed var anfægtet i den version af NT, som dengang var til afprøvning. Erfaringerne herfra kunne tyde på, at det netop kunne være en fordel, at administrationen af det adaptive princip ikke tydeliggør et sammenhængende hele alt for meget. Derved undgås, at den latente helhed, som *er* til stede som konsekvens af opgavehomogeniteten, kommer til at optræde som en platform for elevens tænkning a la: *"Når jeg har svaret sådan her på det ene spørgsmål, så må jeg svare sådan her på det andet"*(som er en tydelig illustration af *manglende* lokal uafhængighed).

Det er en konsekvens af konstruktionen af NT, at det didaktiske "spillerum" begrænses ved rapporteringen af en elevs resultat i forhold til de muligheder, som tradition lineær testning tillader, fordi NT alene rapporterer tre adskilte resultater, fra rent normative målinger, ét mål fra hvert af de tre faglige/didaktiske underdomæner. Tre profilområder dækker et mindre område sammenlignet med de facetter, der kan bygges ind i egentlige diagnostiske prøver og leverer dermed ikke i tilstrækkeligt omfang et detaljeret billede af eleven. Dertil kan tilføjes, at indplaceringen i de kvalitativt beskrevne *niveauplaceringer 1-5* "over gennemsnittet - "under gennemsnittet" ud fra det kvantitative mål på percentilskalaen er et groft informationsgrundlag til forældrene sammenlignet med de tilbagemeldinger via karakterskalaer, som forældre og elever er vant til. Der er tydelig forskel på de krav til "detalje", som lærere og forældre stiller til tilbagemeldingerne af NT-resultaterne. Det er også en konsekvens af selve konstruktionen og implementeringen af NT, at lærere ikke behøver at bruge så megen tid til at fortolke og udlægge besvarelsen af NT sammenlignet med klassiske test. Fx udfordres læreren ikke af den situation, hvor nogle elever kan klare nogle bestemte opgaver (i en lineær test), mens andre elever ikke kan. Problemet falder på en måde væk under det adaptive system, hvor der netop stilles *forskellige* opgaver til forskellige elever. Der er to situationer under en elevtest ved hjælp af NT, som kalder på reaktioner fra lærerne: At skærmen forbliver "gul" – dvs. beregningsrutinerne har svært ved at bestemme elevens niveau – og, som den anden situation, fornemmelsen af at eleven under testen er mere optaget af at tænke "taktisk" i forhold til den måde det adaptive princip virker end at prøve at løse de stillede opgaver. Den sidste situation kan udmærket være en del af den første. Under alle omstændigheder fører den først beskrevne situation til, at testningen på et tidspunkt afbrydes, og man konstaterer, at eleven *ikke* er blevet båret frem til en præcis bedømmelse via det adaptive princip. Denne situation er forudset i udbuddet af opgaven med at konstruere selve NT, og det blev fastlagt som betingelse, at den opstår i mindre end 10 pct. af tilfældene. Man kan imidlertid acceptere den "gule" bedømmelse som ligeværdig med en lineær test indeholdende præcist de samme opgaver.

Der findes elever, hvor indførelsen af det adaptive princip fremkalder fornemmelsen af "spil", som de genkender fra utallige computerbaserede udfordringer men omfanget kendes ikke præcist[16]

Hensigtsmæssigheden af formidling af de nationale test

Hensigtsmæssigheden af den formidling af resultater fra NT, som finder sted, skal belyses i forhold til de psykometriske forudsætninger, hvorunder resultaterne er skabt. Det er således en forudsætning for vurderinger af hensigtsmæssigheden, at den psykometriske metode (model) netop tilhører den klasse af IRT-modeller, som NT er bygget på, som fx at medtage et specifikt mål for elevdygtigheden[17]. Med denne indskrænkning af mulighederne for en vurdering af hensigtsmæs-

[15] *Detecting person misfit in adaptive testing using statistical process control techniques* (Edith M. L. A. van Krimpen-Stoop and Rob R. Meijer).
Evaluation parameters for computer-adaptive testing http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8535.2005.00525.x/full
De nationale it-baserede test i folkeskolen. Rapport fra REVIEW-panelet, Devo Team Consulting, 2007

10 De nationale it-baserede test i folkeskolen. Rapport fra REVIEW-panelet, Devo Team Consulting, 2007

[17] Klassiske psykometriske metoder opererer med elevvariationer, som beskrives via fælles normalfordelinger, og undgår derfor at operere med eleven som en fast, systematisk størrelse.

sighed[18] inddrages det adaptive princip ved sammenligninger mellem traditionelle lineære test og NT.

Det er veldokumenteret[19], at resultater fra prøver og test i forbindelse med pædagogiske evalueringer i almindelighed ikke bliver benyttet i det omfang, man kunne forvente. Dette gælder traditionelle lineære test, mens evalueringen af NT skal belyse, hvordan det forholder sig med lærernes brug af NT. Dette faktum vurderes til i høj grad at være uhensigtsmæssigt i forhold til forventninger om, hvad denne formidling bør kunne igangsætte af ændringer i undervisningen og som et andet aspekt skabe grundlag for elevens selvvurdering. Sammenligner man det faglige indhold af et formidlet resultat fra en traditionel lineær test med resultater fra NT, er der indlysende et større ejerskab til traditionelle prøver og test (som er delvist selvfremstillede), og derfor er det for lærerne nemmere at overskue, anvende og fortolke resultatet fra traditionelle lineære test end resultater fra NT. Dette kan til en hvis grad tilskrives det adaptive princip, som for NT's vedkommende anbringer resultaterne på lidt "anonyme" områdeskalaer. Skalaer, som set sammen (de tre profilområder), giver et indtryk af formativ tilbagemelding, men som er mindre præcist relateret til et curriculum, noget som en traditionel lineær test normalt forholder sig tæt til.

I denne forbindelse vil hensigtsmæssigheden ved formidlingen af resultater fra netop adaptive test kunne ses som en balance mellem at anvende testresultaterne formativt og anvende dem som korte summative evalueringer med områdespecifikke normative målinger. Spørgsmålet om at håndtere denne balance for værdien af det hensigtsmæssige skal ses i relation til ønsker om at opnå præcision og effektivitet (antal stillede opgaver); to forhold, som har været den gennemgående drivkraft bag ved de fleste implementeringer af adaptive systemer[20].

Det er en del af NT, at læreren kan supplere elevprofilen over de tre områder med en oversigt over præsenterede opgaver og elevens svar. Der hersker usikkerhed mht. omfanget af lærernes brug af denne facilitet, som ellers kunne give et fingerpeg om, i hvilken udstrækning lærerene med rette kan betragte de summariske tre-profilben-opsummeringer som tilstrækkelige for deres videre undervisning med eleverne.

En afgørende side af et testresultats hensigtsmæssighed er testresultatets *prædiktive validitet*, dvs. evne til at forudsige elevens senere præstationer og "adfærd" (fx gennemførelse af specifikke uddannelser). Resultater fra NT har imidlertid endnu ikke været afprøvet i denne forbindelse, og analysevinklen tages derfor ikke op her. Imidlertid står det klart, at det adaptive princip i NT gør fortolkningen af prædiktionen vanskeligere sammenlignet med situationer, hvor traditionelle lineære (velkendte) testresultater lægges til grund for prædiktionen, fordi grundlaget for prædiktionen enten skal sammenfattes i de tre relativt grove profilben-opgørelser, eller læreren skal have overblik over detaljerede besvarelser af opgaveforløb, der varierer fra elev til elev.

På det praktiske niveau er det måske vigtigere, at eleven har været præsenteret for opgaver, som "passer" til dygtighedsniveauet end en fornemmelse af, at testresultatet har høj prædiktiv validitet?

Måling af faglig progression

Spørgsmålet om at være i stand til at opstille valide mål for faglig progression er dels afhængig af den anvendte psykometriske metode, der ligger til grund for fortolkning og håndtering af elevernes svar, og dels afhængig af psykometriske egenskaber ved opgaverne i itembanken. Begge

[18] Den danske version af NT er unik og findes ikke andre steder i verden

[19] *Pædagogisk Brug af Test – et systematisk review*. Dansk Clearing House, 2009

[20] Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4
Adaptive test – en pædagogisk udfordring og et didaktisk guldkorn http://www.ind.ku.dk/mona/2010/MONA-2010-1-KommentarMetteRoseEriksenLarsPeterBechKjeldsen.pdf/

disse grundforhold ved måling af faglig progression medtager det adaptive princip som nødvendigt element. Det er samtidig en slags "teknisk reparation" på nogle bestemte, isolerede forhold omkring fagligt set meget dygtige og meget svage elever.

Konsekvensen af, at opgaverne i itembanken alle er konstrueret således, at svarene på dem kan beskrives ved hjælp af en IRT-model (Rasch-model), bliver, at elevernes dygtighedsniveau kan beregnes ud fra et vilkårligt delsæt af opgaverne. Dette forhold legitimerer generelt sammenlignelighed af elevresultater opnået via det adaptive princip (forskellige opgaver).

Når en elev gennemgår en (positiv) faglig udvikling, bør eleven stilles over for udfordringer, der passer til de niveauer, eleven gennemlever. Det er her, det adaptive princip i kombination med de nævnte egenskaber ved opgaverne i itembanken (itemhomogenitet) sikrer, at målene for elevens ændrede dygtighed kan placeres på én skala, hvorpå den faglige progression måles som simple forskelle mellem beregnede dygtigheder ved hvert målepunkt.

Ved traditionelle lineære test kan man også foretage måling af faglig progression. Det foregår ved at benytte faste sæt af opgaver, tilpasset det aldersniveau, som eleven befinder sig på, og derefter foretage det, der i litteraturen betegnes som *test equating*, hvor resultaterne fra parallelle test sammenlignes. Især Dansk Psykologisk Forlag har i Danmark stået for materialer og metoder (Rasch-modeller), som sikrer, at der kan "bygges bro" mellem flere parallelle test. For at disse traditionelle test skal kunne levere valide mål af elevernes faglige progression, skal opgaverne under alle omstændigheder være del af samme system af IRT-godkendte udfordringer, fx sådan som NT er bygget op. På dette punkt er der derfor ingen forskel mellem det scenarie, der udspilles ved NT, og ved de traditionelle lineære test. Ved brug af lineære test, som er bundet sammen via test equation, opstår forskellen til NT især ved den måde, fagligt set meget stærke og meget svage elever behandles. Ved de lineære test rammer sådanne elever enten "loftet" eller "gulvet" og unddrager sig (ved maksimal score eller nul-score) et egentligt kvantitativt mål for dygtigheden. For sådanne elever vil det adaptive princip sikre, at målingerne hele tiden foregår "i midten" af deres formåen og dermed sikrer pålidelige mål af elevdygtighederne. Det adaptive princip kommer dermed til at spille en vigtig rolle ved måling af faglig progression, når der ønskes valide beregninger og sammenligninger af elevdygtighederne.

På det praktiske felt har IEA og OECD erfaringer med måling af progression gennem TIMSS og PIRLS undersøgelserne og PISA[21]. Disse undersøgelser er baseret på opgaver (items), som er homogene i samme psykometriske forstand, som er gældende for items i NT (itembanken). Der sikres mulighed for at foretage måling af faglig progression på populationsniveau, ikke individuelt niveau.

Forskellige studier bygget på statistisk simulation har klart demonstreret overlegenheden af test med det adaptive princip implementeret i forhold til traditionelle lineære test[22]. Det kan vises, at der finder en systematisk overestimation af elevens udviklingskurve sted, især når udviklingen "hopper op og ned", dvs. følger en irregulær, ikke-monoton form. Grundlæggende reflekterer denne observation det faktum, at det adaptive princip netop "opdager" en udvikling – og tilpasser testen derefter – når elevens udviklingskurve viser uforudsigelige bevægelser. Evnen til korrekt at måle faglig progression har stor betydning, hvor fx krav om inklusion netop kan afstedkomme ikke-monotone udviklingskurver for nogle af eleverne. I sådanne tilfælde måles den faglige progression bedre med det adaptive princip sammenlignet med traditionelle lineære test.

[21] TIMSS med matematik og naturfag, PIRLS læseundersøgelse, begge curriculum orienteret, PISA matematik, naturfag og læsning, literacy orienteret

[22] Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4

Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement
http://iacat.org/jcat/index.php/jcat/article/view/16/3

Det adaptive princip som grundlag for ny viden

For at vurdere fremkomst af *ny viden* i forbindelse med afholdelsen af en test, skal det indledningsvis slås fast, at den viden, der er tale om, alene identificeres gennem elevernes svar på de stillede opgaver. Hvis der optræder ny viden i denne forbindelse, kan det derfor henføres til hovedsageligt to områder. Det ene område beskrives ved, at eleven kan svare (korrekt) på spørgsmål, som man ikke forventede, eleven overhovedet kunne tage stilling til, enten fordi eleven skønnes at have begrænsede kundskaber, eller fordi opgaven ligger uden for det curriculum, som eleverne skal testes inden for, men selv om eleven på den måde involveres i ny viden, har det adaptive princip i denne sammenhæng ikke noget at gøre med fremskaffelse af ny viden.

Når elever. der normalt præsterer meget lavt eller højt, møder udfordringerne i en traditionel lineær test, ender resultatet ofte med, at eleven scorer nul rigtige eller maksimum rigtige besvarelser. Den "viden" om eleven, der hermed etableres, er en relativ unuanceret erkendelse af, at eleven enten er svagere, end den letteste opgave stiller af krav, eller at eleven er dygtigere, end den sværeste opgave stiller af krav. I begge tilfælde en evaluering, som ikke er båret af et egentligt kvantitativt mål[23] af elevens dygtighedsniveau. I dette tilfælde vil det adaptive princip fremskaffe *ny kvantitativ viden* om elevens niveau, fordi eleverne under dette princip konstant præsenteres for nye opgaver, der slutteligt sikrer ca. 50 pct. chance for at svare rigtigt - dvs. uden mulighed for at ende med en nul- eller maksimum score.

Det andet område drejer sig om den kendsgerning, at eleven "trækkes igennem" en konkret serie af opgaver/spørgsmål, der, analyseret som *kombinationer* af svar, evt. peger på ny viden. Dette kan være tilfældet som element af formativ evaluering, netop gennem analyser af *kombinationer af svar* på flere opgaver/spørgsmål. I NT præsteres ny viden om eleven i forhold til en *normativt orienteret* konklusion, der blot markerer, at eleven enten kan eller ikke kan besvare et eller flere konkrete spørgsmål.

NT's diagnostiske grundelement består i dannelsen og formidlingen af information fra de tre profilben, *elevprofilen,* som refererer til tre faglige delområder. Ny viden holdes derfor inden for rammerne af disse elevprofiler. Antallet af profilområder hænger naturligt sammen med dannelsen af ny, differentieret viden, men selve beslutningen om at fremstille og præsentere eleven for områdespecifikke faglige elementer har ikke noget at gøre med det adaptive princip og vil heller ikke have det, selv om antallet af profilområder ændres.

Derfor reduceres spørgsmålet om eventuel ny viden i forbindelse med det adaptive princip til et spørgsmål om fremskaffelse af ny viden via *svar på kombinationer af opgaver* og om dette fænomen stimuleres eller ej ved anvendelsen af det adaptive princip.

Det adaptive princip anvendes som grundlag for udvælgelsen af "næste opgave3 inden for hvert profilområde. Udvælgelsen sker tilfældigt blandt de opgaver, som stilles til rådighed for det adaptive princip, dvs. opgaver der er ens mht. sværhedsgrad og reference til læringsmål. Når man derfor ser på rækkefølger af opgaver med deres besvarelser, som er genereret under det adaptive princip, er det eneste, der "binder" opgaverne sammen en systematik mht. sværhedsgraderne, ikke indholdet. Dette kan opfattes som en begrænsning af mulighederne for at foretage formative evalueringer inden for rækken af opgaver i et givet profilområde, sammenlignet med de muligheder, traditionelle lineære test giver. Eller som et citat fastslår[24]: "*Test er - som evalueringsmetode - bedst egnet til at måle elevernes evne til at huske fakta. De er ikke så gode til at teste deres evne til at tænke i sammenhænge og deres evne til at huske og lære sammen med andre. Det betyder, at en meget vigtig del af det, skolerne lærer vores børn, slet ikke bliver testet".* Ved de traditionelle lineære test kan man på forhånd tilrettelægge opgaverne i en rækkefølge, som netop tillader sådanne formative evalueringsmuligheder baseret på sammenhæng skabt over flere opgaver.

[23] IRT-modeller tilordner af matematiske grunde ikke latente dygtighedsværdier til elever med nul- eller maksimum score.

[24] Kousholt K.: *De nationale tests er ikke objektive*

Teknisk og økonomisk egnethed af det adaptive princip

Den tekniske egnethed af det adaptive princip er allerede omtalt tidligere i dette notat i forbindelse med aspekter som "fortolkning af elevsvar i relation til IRT-modeller", "antal stillede opgaver", "præcision ved bestemmelsen af elevdygtighed" og "egnethed til at teste eleven i forskellige læringsmål". Alle steder i denne vurdering er egnetheden blevet belyst via sammenligninger med testscenarier, hvor traditionelle lineære test benyttes i stedet for NT.

Når egnethed af det adaptive princip skal anskues fra en økonomisk synsvinkel, kunne man anlægge den betragtning, at en simpel *cost-benefit analyse* bør kunne belyse økonomiske fordele og ulemper forbundet med det adaptive princips anvendelse under hvert af punkterne 1: *Teknisk grundlag for det adaptive princip* til pkt. 7: *Det adaptive princip som grundlag for ny viden* - set i relation til betingelserne under traditionelle lineære test. Det er af mange indlysende grunde ikke muligt. Det er fx ikke klart, hvilken samlet økonomisk fordel anvendelsen af det adaptive princip har i forhold til den tid, der afsættes til NT. Den "ledige" tid, som opstår for de elever, hvis dygtighed hurtigst fastlægges, kan ikke umiddelbart kapitaliseres til at udgøre en resurse for andre undervisningsrelaterede aktiviteter, fordi der er pædagogiske udfordringer ved at styre en klasse, hvor nogle elever bliver færdige før andre og måske kræver opmærksomhed, mens læreren sørger for, at resten af eleverne arbejder under rolige testforhold frem til grøn skærm. Tilsvarende er det ikke muligt at pege på præcise økonomiske konsekvenser af, at ekstremt svage eller stærke elever kan tilskrives reelle kvantitative mål for deres dygtighed under det adaptive princip, sammenlignet med traditionelle lineære test, hvor disse to typer elever alene klassificeres i grove kategorier uden numerisk angivelse af præstationen[25].

Den økonomiske egnethed belyses ved at pege på og beskrive nogle aktivitets- og omkostningsscenarier, som er forbundet med gennemførelse af traditionelle lineære test og ved test, som er udviklet under adaptive principper – underforstået, som del af det eksisterende NT. Der er tale om at beskrive faser af produktion og analyse af opgaver til NT, delvist ud fra internationale erfaringer og delvist ud fra den kendsgerning, at opgaverne er statistisk relateret til IRT, den psykometriske metode, som har dannet grundlag for NT.

De elementer, der rent arbejdsmæssigt indgår i disse faser, strækker sig fra opgaveudvikling over afprøvning af opgaver på skolerne til efterfølgende Rasch-analyser til opgaverne, som, for NT's vedkommende, efterfølgende skal placeres i opgavebanken. Det skal i denne forbindelse markeres, at NT's første udviklingstrin, indeholdende konstruktion af itembanken, er en aktivitet, som ligger forud for de scenarier, som aktuelt beskrives. De beskrevne faser omfatter derfor alene aktiviteter, som gælder fra skoleåret 2012-2013, hvor modellen var mere i drift, om end der stadig har været visse udviklings- og indkøringsproblemer.

Trods manglende kendskab til præcise omkostninger ved enkeltdele af vedligeholdelsen af NT og af de sædvanlige lineære test, er det en klar, samlet vurdering, at vedligeholdelsen af NT kræver flere resurser end udvikling af fra-gang-til-gang lineære test. Dette gælder også, selv om den fase, hvori man afprøver nye opgaveforslag til NT, kan smidiggøres ved den beskrevne procedure (kaldet "skyggeteknik"), hvorunder nye opgaver "smugles" ind blandt de eksisterende (og velafprøvede) opgaver fra itembanken, uden at tælle med i den endelige estimation af elevdygtigheden. Dette er blot en fiks måde at gennemføre de Field Trial operationer, som er kendt fra PISA og IEA's test (TIMSS og PIRLS), men som i takt med udvidelsen af itembanken bliver mere og mere omfattende, fordi det adaptive princip sætter stærke krav til nye opgaver, som skal indlemmes i itembanken. Den itemhomogenitet, som fordres opfyldt inden for IRT (Rasch)-modellens rammer kan fx ikke valideres ved at tage *samtlige* indgående opgaver (fra et område/profildomæne) i itembanken på én gang. Der findes ikke simple løsninger på en praktisk tilrettelæggelse af indtag og afprøvning af nye opgaver med respekt for det faktum, at en hvilken som helst opgave skal kunne indgå "homogent" i samtlige adaptivt inspirerede opgaveforløb (fra et givet profildomæne). Hvis man slækker på disse homogenitetskrav umuliggør man valide sammenligninger af elevernes præstationer i lys af IRT. Den groveste forenkling af det samlede system med vidtgående, positive konsekvenser for NT-økonomien består i at opgive adaptiviteten og i princippet lade elever fra samme veldefinerede *populationer* (klasser, skoler, kommuner, re-

[25] Manglende mulighed for at tildele elever med nul-score og maksimum score et estimat

gioner eller lignende) *få samme opgaver*, gerne udvalgt fra en opgavebank, der løbende er opbygget af items, som er IRT-homogene.

Litteratur som helt eller delvist lægges til grund for vurderingen

1. *Item selection and ability estimation in adaptive testing* (Wim J. van der Linden and Peter J. PashIey).

2. *Constrained adaptive testing with shadow tests* (Wim J. van der Linden).

3. *Principles of multidimensional adaptive testing* (Daniel O. Segall).

4. *Multidimensional adaptive testing with Kullback-Liebler information item selection* (Wim J. van der Linden and Joris Mulder).

5. *Sequencing an adaptive test battery* (Wim J. van der Linden).

6. *Adaptive tests for measuring anxiety and depression* (Otto B. Walter).

7. *MATHCAT: A flexible testing system in mathematics education for adults* (Alfred J. Verschoor and Gerard J. J. M. Straetmans).

8. *Implementing the Graduate Management admission test computerized adaptive test* (Lawrence M. Rudner).

9. *Designing and implementing a multistage adaptive test*: The uniform CPA exam (Gerald J. Melican, Krista Breithaupt, and Yanwei Zhang).

10. *A Japanese adaptive test of English as a foreign language: Developmental and operational aspects* (Yasuko Nogami and Norio Hayashi).

11. *Innovative items for computerized testing* (Cynthia G. Parshall, J. Christine Harmes, Tim Davey, and Peter J. Pashley).

12. *Designing item pools for adaptive testing* (Bernard P. Veldkamp and Wim J. van der Linden).

13. *Assembling an inventory of multistage adaptive testing systems* (Krista Breithaupt, Adelaide A. Ariel, and Donovan R. Hare).

14. *Item parameter estimation and item fit analysis* (Cees A.W. GIas).

15. *Estimation of the parameters in an item-cloning model for adaptive testing* (Cees A. W. GIas, Wim J. van der Linden, and Hanneke Geerlings).

16. *Detecting person misfit in adaptive testing using statistical process control techniques* (Edith M. L. A. van Krimpen-Stoop and Rob R. Meijer).

17. *The assessment of differential item functioning in computer adaptive tests* (Rebecca Zawick).

18. *Multi-stage testing: Issues, designs, and research* (April Zenisky, Ronald K. Hambleton, and Richard M. Luecht).

19. *Three-category adaptive classification testing* (Theo J.H.M. Eggen).

20. *Testlet-based adaptive mastery testing* (Hans J. Vos and Cees A. W. Glas).

21. *Adaptive mastering testing using a multidimensional IRT model* (Cees A. W. Glas and Hans J. Vos).

22. *Adaptive tests for measuring anxiety and depression* (Otto B. Walter). http://onlinelibrary.wiley.com/doi/10.1002/mpr.274/pdf) Marts 2009

23. *Using Computerized Adaptive Testing to Evaluate Nurse Competence for Licensure*: (http://download.springer.com/static/pdf/513/art%253A10.1023%252FA%253A1009866321381.pdf?auth66=1360742281_617bd5d641f0b9380ca866dbe2a3473a&ext=.pdf)

24. *Some considerations related to the use of adaptive testing for the common core assessments*
http://www.pearsonassessments.com/NR/rdonlyres/76E049D3-2226-472C-96C2-226AE2D9E396/0/TMRS_WP_CAT_Paper_common_core_110310.pdf
http://link.springer.com/article/10.1023%2FA%3A1018420418455?LI=true

25. *Measuring Individual Growth With Conventional and Adaptive Tests*
https://www.assess.com/docs/Weiss-Measuring-Individual-Change.PDF

26. *De nationale tests er ikke objektive 20 dec 2011* (Kristine Kousholts afhandling)
http://videnskab.dk/kultur-samfund/de-nationale-test-er-ikke-objektive

27. *Evaluation parameters for computer-adaptive testing*
http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8535.2005.00525.x/full
28. *Wiley online library*
http://onlinelibrary.wiley.com/advanced/search/results

29. *Evaluation of Computer Adaptive Testing Systems*
http://www.igi-global.com/viewtitlesample.aspx?id=2979&ptid=34720&t=evaluation+of+computer+adaptive+testing+systems

30. *Assessing self-care and social function using a computer adaptive testing version of the Pediatric Evaluation of Disability Inventory Accepted for Publication, Archives of Physical Medicine and Rehabilitation*
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666276/

31. *Adaptive test – en pædagogisk udfordring og et didaktisk guldkorn*
http://www.ind.ku.dk/mona/2010/MONA-2010-1-KommentarMetteRoseEriksenLarsPeterBechKjeldsen.pdf/

32. *De adaptive test er bedst*
http://www.skoleraadet.dk/sitecore/content/Skoleraadet/Aktiviteter/Seminarer/Seminarer%202007/~/media/Skoleraadet/Aktiviteter/Seminarer/Seminarer%202007/Foraar/julius%20bj%C3%B8rnsson%20referat.ashx

33. *Common Adaptive Tests to Address Special Needs Questions raised about adaptive assessments*
http://www.edweek.org/dd/articles/2012/10/17/01adaptside.h06.html

34. *The Transition to Computer-Based Assessment New Approaches to Skills Assessment and Implications for Large-scale Testing*
http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/8713/1/reqno_jrc49408_final_report_new%281%29%5B1%5D.pdf?st=3623

35. *Læseforsker skyder modellen bag de nationale test ned*
http://www.folkeskolen.dk/508727/laeseforsker-skyder-modellen-bag-de-nationale-test-ned

36. *Rigsrevisor: Nationale test er stadig risikofyldt*
http://www.folkeskolen.dk/507872/rigsrevisor-nationale-test-er-stadig-risikofyldt

37. *International guru: Tag imod de nationale test som en chance*
http://www.folkeskolen.dk/511101/international-guru-tag-imod-de-nationale-test-som-en-chance
Computerized Adaptive Testing. How CAT May Be Utilized in the Next Generation of Assessments. A Report for the North Carolina State Board of Education
http://www.ncpublicschools.org/docs/acre/publications/2010/publications/20100716-01.pdf

38. *Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education*
http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/we04070.pdf

39. *Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement*
http://iacat.org/jcat/index.php/jcat/article/view/16/3

40. *Journal of Computerized Adaptive Testing*
http://iacat.org/jcat/index.php/jcat/issue/current

41. *The design and evaluation of a computerized adaptive test on mobile devices*
http://www.sciencedirect.com/science/article/pii/S0360131506001965

42. *A Practical Computer Adaptive Testing Model for Small-Scale Scenarios*
http://ifets.info/journals/11_3/18.pdf

43. *Practical Implications of Item Response Theory and Computerized Adaptive Testing: A Brief Summary of Ongoing Studies of Widely Used Headache Impact Scales*
http://www.jstor.org/stable/10.2307/3768065

44. Guido Makransky: *Computerized Adaptive Testing in Industrial and Organizational Psychology* Ph.D. thesis, University of Twente The Netherlands 2012 ISBN: 978-90-365-3316-4

45. *De nationale it-baserede test i folkeskolen.* Rapport fra REVIEW-panelet, Devo Team Consulting, 2007

46. *Pædagogisk Brug af Test – et systematisk review*. Dansk Clearing House, 2009

Appendiks: Antal opgaver per profilben

Profilområde	Antal opgaver
010201	208
010202	279
010203	202
010401	226
010402	284
010403	243
010601	264
010602	229
010603	222
010801	162
010802	212
010803	193
020301	231
020302	223
020303	235
020601	334
020602	265
020603	244
030801	247
030802	266
030803	236
040801	220
040802	211
040803	194
050501	191
050502	187
050503	167
050701	223
050702	215
050703	264
060701	215
060702	252
060703	209
070801	219
070802	242
070803	218
Totalt antal opgaver	8232