Virum og Odense, 5. januar 2018

Svend Kreiner, professor emeritus

Jeppe Bundsgaard, professor MSO

Åbent brev til Undervisningsminister Merete Riisager og medlemmerne af Folketingets Undervisningsudvalg

Vi har med interesse fulgt og deltaget i diskussionen om De Nationale Test (DNT), og selvom vi ikke nødvendigvis er enige i alt hvad der siges både for og imod pædagogiske test i al almindelighed og DNT i særdeleshed, er det åbenlyst for os begge, at DNT ikke er blevet det nyttige redskab for alle lærerne i den danske folkeskole, som pædagogiske test kan og bør være.

Det er åbenlyst, at der er meget, der kan og bør forbedres i forbindelse med DNT, og da det nu ser ud som om der er politisk interesse i at gøre noget ved det, vil vi med dette brev give en række indspark til, hvad der kan forbedres.

I en ikke helt systematisk rækkefølge bør følgende ting komme på banen.

Brugen af DNT skal tilrettelægges således at den er nyttig for læreren

De nationale test har fra starten været præsenteret som et pædagogisk redskab. Men der findes ikke overbevisende belæg for, at de faktisk har udviklet sig til et sådant.

Det kan der være flere årsager til. En af de væsentligste er formodentlig, at afprøvningen er tilrettelagt uden respekt for og uden hensyn til, hvornår det vil være nyttigt for læreren at få de oplysninger om eleverne, som testresultaterne kan give. Det forekommer os desuden at der ikke tages tilstrækkeligt hensyn til om testene rent faktisk giver læreren de oplysninger, som læreren har brug for, for at kunne tilrettelægge undervisningen på den bedste måde. Da det, der er nyttigt at vide og hvornår, må forventes at variere fra klasse til klasse, vil vi opfordre til følgende:

- Læreren skal selv bestemme, hvornår der skal testes. Hvis læreren mener, at det er bedre at teste eleverne i oktober og november end sent i skoleåret, fordi tidlig viden giver bedre muligheder for at tilrettelægge undervisningen i forhold til de problemer, som testresultaterne måtte afsløre, skal læreren naturligvis have ret til at træffe den beslutning.
- Læreren skal kunne fravælge irrelevante profilområder så testningen fokuserer på og giver mere sikre resultater om det relevante. Hvis en lærer i 8. klasse fx mener, at det er spild af tid at teste hele klassen i afkodning, skal læreren have ret til at fravælge dette profilområde for nogle eller alle elever, således at tiden i stedet kan bruges til at teste elevernes tekstforståelse. En sådan beslutning vil kunne reducere usikkerheden på målingerne af tekstforståelse til ca. 70 % af den usikkerhed, man har i dag.
- Hvis læreren oplever, at testene ikke bidrager med noget, hun kan bruge til undervisningen – fx på grund af usikkerheden, eller fordi hun ikke forstår testresultaterne – skal hun have mulighed for at fravælge testen og i stedet bruge undervisningstiden på noget mere nyttigt.

Ovennævnte forslag er først og fremmest motiveret af det synspunkt, at anvendelsen af testene skal være så nyttige som muligt for undervisningen, og at tid, der kun fungerer som spildtid for læreren og eleverne, skal reduceres til det mindst mulige. Udover det, vil implementeringen af forslagene kunne betyde, at lærernes engagement i testene forøges, fordi de kommer til at opleve, at de i højere grad er med til at styre, hvorledes testene bruges.

En anden årsag til at den nationale test ikke har udviklet sig til det nyttige pædagogiske redskab, som man havde håbet, er, at det ifølge en undersøgelse af Bundsgaards og Puck (2016) er mindre end 10 % af lærerne, der forstår, hvad resultaterne betyder. Udover at dette tal i sig selv er chokerende lavt, er det naturligvis illusorisk, at lærere, som ikke forstår testresultaterne, kan få noget som helst ud af dem, som de kan bruge i undervisningen. Et redskab forudsætter, at brugerne forstår at bruge det. Det gælder også for pædagogiske test.

Den mangelfulde indsigt i, hvad testresultater betyder, skyldes ikke lærerne. Det skyldes udelukkende, at man ikke har forstået at formidle tingene på en ordentlig måde. Det er ministeriets ansvar, og vi kan kun opfordre til at ministeriet tager problemet alvorligt og gør noget ved det. Testresultaterne *skal* gøres forståelige og meningsfulde for lærerne, hvis lærerne skal kunne bruge dem til noget fornuftigt. Det kan ske gennem

- Kurser.
- Meget bedre forklaringer af hvad testresultaterne betyder.
- Mindre komplicerede præsentationer af resultater (i øjeblikket gengives de samme resultater på mindst 3 forskellige måder).
- Ordentlige kriteriebaserede proficiency scores. Læreren har brug for at vide hvilke dele af stoffet, eleven har store problemer med, hvilke dele eleven kan arbejde med uden uovervindelige problemer, og hvilke dele der ikke længere giver problemer for eleven. Testresultater, der leverer den slags oplysninger, omtales som *proficiency scores*, og det er den form for testresultater, som læreren kan drage nytte af i forbindelse med undervisningen af eleverne og klassen. De testresultater (inkl. de såkaldte kriteriebaserede scores), som DNT leverer, fortæller kun noget om, hvorvidt eleven er dygtig eller mindre dygtig. Det ved læreren i langt de fleste tilfælde allerede, og dermed bidrager DNT ikke med noget, som læreren kan drage nytte af.

Proficiency scores er den bedste måde at formidle testresultater på, hvis resultaterne fra testene skal bidrage til lærernes fagligt-pædagogiske arbejde. Proficiency scores udvikles ved, at faglige eksperter analyserer og beskriver, hvad der kendetegner opgaver på forskellige niveauer, og på den baggrund udarbejder en beskrivelse af normal progression inden for det faglige område. Den enkelte elevs resultat kan så relateres til denne progression, og der kan opnås viden om, hvad eleven har af udfordringer lige nu og skal til at arbejde med. På grund af hemmelighedskræmmeriet omkring opgaverne i de nationale test, er det ikke muligt for os at sige, om opgaverne i nationale test i den nuværende udformning indeholder et tilstrækkeligt udfoldet fagligt indhold, til at det er muligt at konstruere egentlige proficiency scores. Det bør derfor undersøges, om det er muligt, og der bør udvikles nye opgaver, der gør det muligt at konstruere proficiency scores, hvis de eksisterende opgaver ikke er tilstrækkelige til formålet.

Spørgsmålene om DNTs validitet og nytteværdi skal håndteres ordentligt

En lang række fagdidaktiske eksperter og lærere har peget på at DNT måler for snævre dele af fagene og gør det på en for usikker måde.

Derfor bør det dokumenteres, at profilområderne er fagligt set meningsfulde, og at opgaverne dækker alle relevante aspekter af profilområderne (indholdsvaliditet).

Den psykometriske begrebsvaliditet skal forklares og dokumenteres. Internationalt er der tradition for at der udarbejdes tekniske rapporter, der beskriver udviklingsprocessen og de teoretiske baggrunde for test. I forbindelse med DNT findes der intet sådant tilgængeligt forarbejde, og det er derfor ikke muligt for uafhængige forskere at gå arbejdet efter i sømmene.

Selvom målingerne af de forskellige profilområder er psykometrisk valide, er det ikke nødvendigvis givet, at disse profilområder er de mest relevante og nyttige for lærerne i arbejdet med eleverne. Vi opfordrer derfor til, at der lægges op til saglig og faglig diskussion af de valgte profilområder.

For at diskussionen (og i givet fald forsvaret) af DNT skal være mulig, er det nødvendigt, at

- Hemmeligholdelse af indhold, arbejdsprocesser og tekniske forhold mindskes i så høj grad som muligt. Pædagogiske test er andet og mere end standpunktsprøver, der kun har det formål at skille fårene fra bukkene. Der er derfor ingen grund til og heller ikke nogen tradition for at holde opgaverne hemmelige.

- At de personer, der har ansvaret for opgaverne og for definitionen af profilområder, forklarer baggrunden for designet af opgaver og profilområder og forholder sig til kritik.

- At man er parat til at droppe profilområder, som lærere og fagdidaktikere finder irrelevante, og enten erstatter dem med andre eller nøjes med færre, så sikkerheden på resultaterne inden for de profilområder, som er relevante, kan forøges.

- At man forøger indholdsvaliditeten gennem at udvikle flere typer af opgaver (itemtyper). Det er usandsynligt at alle aspekter af et fagligt område kan måles med kun én eller få typer opgaver. For nogle elever kan en opgavetype i sig selv give problemer, og derved vil målingen af elevens dygtighed blive skæv, hvis kunne én type opgaver anvendes.
- Kvaliteten af opgaverne skal kontrolleres løbende, og man bør være parat til både at definere nye profilområder og at udvikle og afprøve nye opgaveformer.

Problemer med elevers og læreres negative oplevelser skal tages alvorligt

Forskning i form af både casestudier (Kousholt 2015a; 2015b) og surveys (Bundsgaard & Puck 2016) og beretninger fra praksis (fx i *Folkeskolen*) har vist, at der er elever, der oplever testsituationen som utryg og alt for svær. Præcis hvor mange elever, der er tale om, er der desværre kun få konkrete oplysninger om. Bundsgaard & Puck rapporterer, at godt 20 % af lærerne oplever at der er en eller flere elever i deres

klasse, der er kede af at blive testet. En endnu ikke publiceret undersøgelse af knap 1100 elevers oplevelser af de nationale test oplyser at 17 % af eleverne synes det er ubehageligt at besvare de nationale test. Et meget forsigtigt gæt er derfor, at det er ca. 20 % af eleverne har dårlige oplevelser med de nationale test. Begge undersøgelser rapporterer samtidig, at knap halvdelen af eleverne har positive testoplevelser, men uanset det kan man argumentere for, at antallet af elever med dårlige testoplevelser er for stort. Da et meget stort antal lærere giver udtryk for, at de betragter testene som en kontrol af deres praksis, og at de derfor spilder megen værdifuld undervisningstid på at træne eleverne til testen (såkaldt *teaching for the test*) vil vi opfordre til:

- at det er lærerens ansvar at vælge sværhedsgraderne på de opgaver, som det adaptive system udvælger til eleverne, således at eleverne har fx 75 % sandsynlighed for at svare korrekt på opgaverne i stedet for 50 % som det er nu. Det vil gøre oplevelsen mindre ubehagelig for eleverne, men vil betyde at der i givet fald skal besvares flere opgaver og bruges længere tid på testen for at opnå samme sikkerhed som nu.
- at lærerne gives mulighed for at vælge, hvilke sværhedsgrader eleverne skal starte med.
- at testenes karakter af *high stakes* (dvs. at lærere og elever potentielt kan imødese sanktioner for dårlige testresultater) fjernes for både elever og lærere. Det kan ske ved at der laves et nationalt gennemsnit på baggrund af en tilfældigt udvalgt gruppe af klasser, i stedet for at alle klasser indgår i de nationale gennemsnit. Antallet af elever, der udtrækkes til dette formål kunne f.eks. svare til det antal elever, som PISA-undersøgelserne betragter som tilstrækkeligt til at vurdere, hvorledes danske elever klarer sig i pædagogiske test.

Testresultater er usikre og uforståelige

Målinger ved hjælp af pædagogiske test er målinger, der altid er behæftet af en vis grad af usikkerhed. Det er tilsyneladende kommet bag på mange, selvom det en mere end 100 år gammel nyhed, og der har været meget kritik af usikkerheden ved resultaterne i nationale test. Selvom en stor del af denne kritik er baseret på en utilstrækkelig viden om, hvad usikkerheden skyldes og betyder, samt af en manglende erkendelse af at almindelige ikke-adaptive test er præget af større usikkerhed end de nationale test, skal problemerne med usikkerheden tages alvorligt.

- Usikkerheden på testresultaterne skal beskrives, så brugerne kan forstå hvad det handler om. Der er allerede taget initiativer i den retning, men det kan gøres endnu bedre. I de situationer, hvor testresultaterne placeres i forhold til et lille antal kategorier, bør oplysningerne om usikkerheden fx suppleres med oplysninger om, hvor stor risiko der er, for at en eleven er placeret i en forkert kategori
- På grund af usikkerheden bør testresultater på individniveau ikke deles med andre end kolleger og evt. skoleledelsen. Heller ikke med forældrene. I forhold til forældrene fungerer testresultaterne kun som meget usikre standpunktsprøver.

- Ministeriets behov for standpunktsprøver skal ikke blandes sammen med og slet ikke dominere lærernes brug af testene.
- Testresultater skal altid ses og vurderes i en kontekst. Det vil sige sammen med alt det, som læreren ved om eleven.

Der er flere måder at reducere usikkerheden i pædagogiske test.

- Den væsentligste faktor til at reducere usikkerheden i en pædagogisk test er, at forøge antallet af opgaver.
- En anden faktor er at reducere antallet af opgaver, der enten er alt for lette eller alt for vanskelige for eleven. Det er på dette punkt - og kun på dette punkt - at adaptive test er bedre end almindelige ikke-adaptive test.
- Udover det er vi bekendt med, at ministeriet er i gang med at undersøge, om der er visse opgavetyper, der bidrager mere til at reducere usikkerheden end andre opgavetyper, og at resultaterne ser lovende ud.
- For at undgå, at testforløbet starter med opgaver, der er alt for lette eller alt for vanskelige for eleven, vil det være en fordel, hvis man lader læreren indplacere den enkelte elevs startniveau således, at de første opgavers sværhedsgrad ligger omkring elevens forventede dygtighedsniveau. Dette har den klare pædagogiske fordel, at læreren efter testen kan se, om den enkelte elev vurderes af DNT til at ligge på det samme niveau, som læreren forventede.
- Ønsket om at teste tre profilområder i hvert testforløb, er en betydelig årsag til den store usikkerhed. Hvis man sætter antallet af profilområder ned (så testen kun skal give et eller to resultater), bliver tid til flere opgaver inden for de andre profilområder.
- En anden mulighed for at forbedre sikkerheden på resultaterne er at forlænge testtiden. Men dette er ikke nødvendigvis en god ide, særligt da DNT opleves som en belastning af mange elever. Ved mere performance-orienterede test kan det dog sagtens lade sig gøre for eleverne at deltage i test i længere perioder.
- Undersøgelser af testenes reliabilitet har vist at nogle profilområder har meget lav test-retest-korrelation. Vi foreslår at fjerne sådanne profilområder, fordi en svag korrelation er et signal om, at usikkerheden på testene er for stor i forhold til spredningen af eleverne.

I pædagogiske test vil der altid være en procentdel af eleverne, som har et atypisk testforløb, hvor i øvrigt svage elever svarer rigtigt på vanskelige opgaver, og dygtige elever svarer forkert på lette. Inden for Rasch-modellen kan man få et statistisk mål for, hvor godt elevens svar ”passer” med det forventede forløb, som kaldes *person fit*. Dette mål kan bruges til at fortælle læreren, at der ikke bør tillægges resultaterne for stor tillid, og til at hindre, at testresultaterne indgår i gennemsnitsberegninger for større grupper af elever.

Yderligere forslag

Hvis testene gøres frivillige, vil det være en naturlig udvikling, at der løbende udvikles test til yderligere områder, således at lærerne får et redskab til at vurdere, hvordan deres elever klarer sig inden for flere væsentlige områder af de ganske omfattende fag, de skal undervise i.

Beregning af resultater for større grupper af elever sker i dag på en teknisk set upræcis måde. Man bør derfor anvende såkaldt plausible værdier i beregningen af gennemsnit osv., så man ikke undervurderer usikkerheden på estimaterne.

Forskning i pædagogiske test bør intensiveres. Det er tankevækkende, at man samtidig med, at man fra centralt hold begyndte at udvikle og deltage i pædagogiske test (PISA og DNT), nedlagde det sektorforskningsinstitut – dvs. Danmarks Pædagogiske Institut – der havde pædagogiske test som særligt ansvarsområde. I forhold til for 25-30 år siden er der næsten ingen forskning inden for dette område i Danmark. Hvis beslutningstagerne mener det alvorligt, at pædagogiske test er nyttige, bør de tage ansvar for at denne forskning genoptages, i stedet for at forlade sig på at konsulentfirmaer som Rambøll, Cowi, Epinon og Damvad nok skal finde nogen, der kan løse opgaverne for dem. Vores oplevelse fra samarbejder med sådanne firmaer er, at de sjældent har været i stand til selv at løfte opgaven på et fagligt acceptabelt niveau og derfor har været afhængige af, om de kunne finde fageksperterne.

Lidt om forfatterne til dette åbne brev

Svend Kreiner er professor emeritus med speciale i statistik og psykometri og har arbejdet med udvikling of afprøvning af pædagogiske test i 49 år. Han var konsulent for ministeriet ved udviklingen af DNT og skrev i den forbindelse flere af de baggrundspapirer, som lå til grund for udvikling af testens statistiske algoritmer. Han har desuden gennemført flere undersøgelser af testens validitet og forbindelse til fx PISA. Kreiner har desuden igangsat en international diskussion af, hvorvidt man på baggrund af PISA-undersøgelsen kan rangordne lande i forhold til deres resultater i PISA.

Kontakt: Tlf.: 26 36 52 15. E-mail: svend.kreiner@mail.tele.dk

Jeppe Bundsgaard er professor MSO med speciale i fagdidaktik og it. Han har deltaget som National Research Coordinator i *International Computer and Information Literacy Study*, han er international ekspert i forbindelse med den nye test af ICT literacy i PISA 2021, og han har deltaget i udvikling af en række innovative test af de såkaldte 21. århundredes kompetencer. Han har sammen med Morten Rasmus Puck gennemført en undersøgelse af danske læreres og skoleledere praksis med, holdninger til og viden om nationale test. Han har desuden undersøgt og kritiseret nationale test fra et fagdidaktisk perspektiv.

Kontakt: Tlf.: 31 19 26 07. E-mail: jebu@edu.au.dk.

Referencer

Bundsgaard, J., & Puck, M. R. (2016). *Nationale test: danske lærere og skolelederes brug, holdninger og viden*. København: DPU, Aarhus Universitet.

Kousholt, K. (2015a). Børns gætterier ved nationale test. CEPRA-striben. Tidsskrift for evaluering i praksis, (18), 46-57.

Kousholt, K. (2015b). *Børn som deltagere i social testpraksis*. Paedagogisk Psykologisk Tidsskrift, 52(3), 63-85.