At trykke på "statistikknappen" er ikke tilstrækkeligt

Peter Allerup, Danmarks Pædagogiske Universitet

Kommentar til kommentaren "De digitale prøver har høj kvalitet" i MONA, 2006(3).

Keld Nørgaard har sammen med Lise Steinmüller og Michael Lund-Larsen svaret på en kritik af de nye digitale prøver i naturfag som blev fremført af Eigil Larsen i MONA, 2006(2). Eigil Larsens kritik falder på flere forskellige planer, hvor de rent biologifaglige aspekter udgør en stor del mens mere teknisk prægede kritikpunkter udgør en anden del, der igen kan opdeles i punkter der har med rent teknologiske forhold at gøre (edb-maskiners standard, adgang til at afvikle webbaserede aktiviteter på skolerne mv.), og punkter som har med prøvernes evne til at fungere som måleinstrumenter af elevfærdigheder i biologi at gøre. Det sidste kan man kort referere til som edb-prøvernes psykometriske egenskaber. Disse blev sat under kritik og forsøgt tilbagevist af Keld Nørgaard et al. i MONA, 2006(3).

Spørgsmålet er så om deres tilbagevisning er fuldt dækkende i forhold til den fremførte kritik. Det synes jeg ikke at den er, og derfor vil jeg kommentere et par af de punkter som har med det psykometriske at gøre.

Det er et skridt i den rigtige retning i forhold til "normal" praksis i forbindelse med udarbejdelsen af 9.-klasses-afgangsprøver at der fra UVM's side stilles krav om at besvarelserne af opgaverne skal kunne beskrives ved hjælp af en Rasch-model – at opgaverne er såkaldt *Rasch-homogene* – uanset om opgaverne laves på gammeldags facon med papir og blyant eller via edb. Hermed indføres der psykometriske krav til opgaverne om at producere bestemte typer af besvarelser som kort beskrevet bl.a. går ud på:

- at eleverne skal kunne sammenlignes ved *alene* at se på antallet af rigtige besvarelser
- at eleverne skal kunne sammenlignes uanset hvilke delopgaver der lægges til grund, og uanset hvilke elever der i øvrigt indgår i sammenligningerne.

Det første krav er ofte ikke opfyldt i praksis fordi der er specifikke opgaver der favoriserer det ene køn eller bestemte elevgrupper med speciel undervisningsmæssig baggrund (lærebøger), eller som virker "skævt" fx fordi dygtige elever har for få rigtige (eller svage elever har for mange); man taler om at opgaverne er *Rasch-inhomogene*. Det andet krav er uomgængeligt nødvendigt i forbindelse med fx PISA og andre internationale undersøgelser samt de snart etablerede nationale test (foråret 2007) hvor sammenligningerne mellem elever baseres på svar fra forskellige opgavehefter, eller ved de kommende test til foråret hvor testen, ved hjælp af en række opgaver, tilpasses den enkelte elev og derfor er unik for netop denne elev (adaptive test).

I praksis sikrer man sig at opgaverne er Rasch-homogene ved først at afprøve en række testopgaver og så på grundlag af det empiriske datamateriale som udgøres af prøveelevernes svar, foretage en statistisk analyse ud fra Rasch-modellen for at se om de to krav opfyldes. Kontrollen af om Rasch-modellen er velegnet som beskrivelse, sker i et tæt parløb med de fagpersoner der har fremstillet prøveopgaverne, fordi det kræver speciel teknisk indsigt at afgøre om Rasch-modellen er gyldig. På dette punkt svigter Keld Nørgaard et al.s tilbagevisning af kritikken fordi deres fremstilling af *hvorledes* kontrollen kan gennemføres ved hjælp af numeriske "Chi-Square Fit Statistics", er forkert. Desuden tilskrives disse teststørrelser fejlagtigt egenskaber egnet til afdækning af "gætning". Deres tilbagevisning af at der "gættes", er derfor helt utilstrækkelig. Set fra et professionelt statistisk synspunkt mangler der dokumentation for den påståede Rasch-homogenitet belyst gennem analyser af ICC-kurverne (Item Characteristic Curves), som er stedet hvor både "fit" og "gætning" normalt diskuteres. Det er desværre ved at være en udbredt forestilling også inden for området IRT (Item Response Theory, som har Rasch-modellen som specialtilfælde) at hvis man bare køber et edb-program (i dette tilfælde WINSTEP fra University of Chicago), så behøver man ikke vide ret meget om matematisk statistik, men kan nøjes med at trykke på edb-knapperne og derefter læse de vigtigste linjer ud for "p-værdierne" der fortolkes som "o.k." eller "signifikant afvigelse" afhængig af om de er større eller mindre end 0,05.

Man opdager som professionelt arbejdende statistiker ret hurtigt at når man har 33.500 elevbesvarelser i et datamateriale af denne type, så bliver "alt" signifikant. Vurderet strikt ud fra sædvanlige p-værdi-vurderinger forkaster man derfor "alt" vedrørende Rasch-homogenitet! Dvs. 33.500 besvarelser er "nok" til at fortælle at selv den mest sofistikerede model kommer til kort som beskrivelse over for så massiv en "virkelighed". Godt for det! Det er derfor omvendt et dårligt tegn for Keld Nørgaard et al.s analyser at de kun forkaster 6 % af delopgaverne som uegnede. Det er faktisk ret umuligt at forestille sig som resultat af en korrekt gennemført statistisk analyse, og den lave procent har intet at gøre med at forarbejdet er udført af fagligt sikre lærere, sådan som Keld Nørgaard et al. fremfører! Det faglige islæt og relationen til opstillede

trinmål har intet med ønskede egenskaber i psykometri at gøre idet psykometrien i stedet for "indhold" handler om rent teknisk "administrative" relationer delopgaverne imellem. Var det så simpelt som Keld Nørgaard et al. siger, havde PISA's og IEA's internationale undersøgelser nok for længe siden lært "trickene", så de kunne undgå at kassere ca. 50 % af de indledende delopgaver (items). I dette tilfælde argumenteres der for prøvernes kvalitet på en måde som ikke er dækket af den matematiske baggrund for teststørrelserne og hvordan de skal fortolkes.

Samme fornemmelse af manglende indsigt i det matematiske grundlag får man når der skrives om målefejlen (measurement error) som noget der øjensynligt kan komme ned på et teoretisk minimum afhængig af hvor "god" prøven er – de burde vide at det er antallet af opgaver der er hovedårsag til store eller små værdier af målefejlen. Det fremførte tal, "0,3 logit", er derfor ikke et argument for prøvens kvalitet. Men det lyder godt – som når min bilmekaniker siger at "succesionsventilen i din controler står 3 µ for højt".

Og nej, Rasch-modellen sætter ikke, som Keld Nørgaard et al. påstår, mål for elevpræstationer og opgavesværheder ind på én fælles skala. Det er en misforstået fortolkning af de såkaldte "item maps" hvor to grundlæggende forskellige skalaer – én for elever og én for opgaver – af praktisk illustrative grunde præsenteres på én skala. Forfatterne glemmer at beskrive at der findes bestemte matematiske forudsætninger for den fælles indplacering – den fælles skala er *ikke* uden at man indtænker disse forudsætninger et sted hvor man kan diskutere om opgaverne passer til elevers fordeling af "dygtigheder".

Idéen om én fælles skala er da øjensynligt også glemt når den sædvanlige frugt af tanken om én skala skal høstes, for forfatterne benytter straks efter *to* skalaer til argumentet om at "opgaverne passer til eleverne". Men nej, det er *ikke* et argument for prøvens kvalitet at fordelingen af elever der besvarer den samlede opgave korrekt på forskellige procenter (fra 0 % til 100 %), er normalfordelt. Faktisk bør en dokumentation for en hvilken som helst fordeling ske ved hjælp af såkaldte (latente) Rasch-scores, idet man bør vide at figur 2's fordeling af procent-rigtige aldrig kan være normalfordelt!

Og nej, det er ikke nødvendigvis noget kvalitetsmærke at opgaverne har sværhedsgrader der er jævnt fordelt over hele aksen, sådan som det er forsøgt illustreret i første figur. Det svarer til at have en målepind med delemærker sat ind overalt på pinden med konstant afstand. Hvis man ønsker at se forskel på elever der bunkes sammen på midten i én grå masse, omkring 64 % rigtige (jf. figur 2), så var det måske en bedre idé at bruge kræfterne på at *samle* delemærkerne (itemsværheder) i et felt så man *kan* se forskel på eleverne i midten.