Kort uttalelse om rapporten «Undersøgelse af De Nationale Tests måleegenskaber» i lys av kritikken som kom fram i kronikk i Politiken i mai 2019

Rolf V. Olsen, professor, Centre for Educational Measurement, Universitetet i Oslo

Utgangspunktet for denne korte uttalelsen er at jeg var reviewer av denne rapportens førsteutkast. Jeg hadde i min review angitt ganske mange små forbedringspunkter (knyttet til hvordan dette noe komplekse stoffet kommuniseres). Mine kommentarer har i stor grad blitt fulgt opp i den endelige rapporten. Jeg har ikke gjort en review av koder og datasett, men lagt til grunn at analysene er gjennomført som beskrevet.

Grunnleggende sett var min review positiv: de data og metoder som er anvendt er relevante for å belyse de spørsmålene som rapporten stiller. Konklusjonene som trekkes i kap. 4.4 er nøkterne og godt begrunnet i resultatene som presenteres. Når det gjelder innholdet i kapittelet med diskusjoner, har jeg i min review sagt at alle videre implikasjoner (utover de nøkterne konklusjonene gitt i 4.4.), får stå for forfatternes regning. Men det er naturlig at forfatterne i en slik rapport diskuterer mulige implikasjoner, og gitt resultatene av analysene, framstår de videre implikasjonene som drøftes i kapittel 6, som rimelige. La meg tilføye at i analyser av gjennomsnittsskårer for grupper av elever på 100 eller mer, så vil nok de fleste hovedkonklusjoner fra tidligere publisert forskning, være robuste overfor de problemene som identifiseres av Bundsgaard og Kreiner.

Jeg har også lest kronikken i Politiken fra en gruppe forskere, i all hovedsak innen økonomi. Jeg ser ikke at denne kronikken problematiserer eller drøfter hovedkonklusjonene i rapporten i vesentlig grad. Utsagn som sier at «alle tester har feil», oppfatter jeg som stråmenn i denne debatten. Det er selvsagtheter som ikke berører det mer interessante: noen tester har større feilmarginer enn andre. Dette drøfter jo også Bundsgaard og Kreiner godt. Den adaptive testalgoritmen opererer i utgangspunktet med en, etter mitt syn, for stor toleranse for feil (<0.55). I tillegg får en stor andel elever feil som er betydelig større enn dette. Til sammen tilsier dette at prøveresultatene ikke kan benyttes for tolkninger av individuelle elevers skårer. Det underliggende hovedproblemet er at testen gir tre skårer til hver enkelt elev basert på et for lavt antall oppgaver. Til sammenlikning kan det nevnes at for de norske nasjonale prøvene (ikke adaptive) er det et absolutt krav om reliabilitet større enn 0.85, og i de fleste tilfellene er reliabiliteten av prøvene i Norge større enn 0.90.

I kronikken i Politiken synes det som om manglende reliabilitet ikke er en interessant problemstilling. Og gitt at man kun skal se på resultater for store grupper av elever, så er dette i ren statistisk forstand forståelig. Men forfatterne ser bort fra at så store målefeil er en trussel mot testens validitet. Og ikke minst gir så store målefeil, som drøftet ovenfor, et stort problem for læreres bruk av resultater. Og nettopp dette siste er et eksplisitt uttrykt formål for de nasjonale test i Danmark.