Aben der nægtede at dø. Multiple choice-prøver og korrektion for gætteri

Lotte Dyhrberg O'Neill,
Universitetspædagogik,
Institut for Kulturvidenskaber, SDU

Kommentar til Sunde & Sunde: "Den smarte abe: betydning af og korrektion for gætning ved karaktergivning i multiple choice-tests", MONA, 2016-4.

Den smarte abe har undersøgt emnet tilfældigt gætteri i multiple choice-tests (MCT) med udgangspunkt i simulerede data for folkeskolens afgangsprøve i biologi fra 2015. Formålet med artiklen var at "kvantificere og illustrere betydningen af tilfældig gætning for sandsynligheden for opnåelse af de forskellige karakterer samt foreslå praktiske løsninger til hvorledes man kan tage højde for dette i karakterudmålingen i forbindelse med kriteriebaseret karaktergivning". Resultaterne viste at det i praksis var umuligt for eksaminanderne at opnå karakteren -3 medmindre de anstrengte sig for at svare forkert, og at selv en fingernem og ihærdig abe ville kunne bestå prøven ved cirka hvert 18. eksamensforsøg. Det sidste ville nok koste nogle bananer. Forfatterne understregede at *de* ikke ville forholde sig til andre relevante aspekter end de matematiske og statistiske aspekter relateret til gætteri i multiple choice-tests (MCT). Tillad mig derfor at gøre netop dette i denne kommentar.

Først synes jeg at det er på sin plads at påpege hvilken større teoretisk ramme diskussionen om gætteri indgår i. Det rette bagtæppe for diskussionen er moderne validitetsteori (Messick, 1989; Kane, 2006; Standards for Educational and Psychological Testing, 2014). Store anerkendte organisationer som American Educational Research Association (AERA), American Psychogical Association (APA) og National Council on Measurement in Education (NCME) har defineret validitet på følgende måde (American Educational Research Association, American Psychological Association, National Council of Measurement in Education, 2014):

> *"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretation."*

Der er mange typer af validitetsevidens som bør vurderes samlet. Overordnet bør der være evidens som støtter svarprocessen, generaliserbarheden, muligheden for at ekstrapolere ud fra scorerne samt de beslutninger der tages på baggrund af scorerne og deres konsekvenser (Kane, 2006). Det påhviler de testansvarlige dels at formulere åbent og tilgængeligt de eksisterende underliggende antagelser og fortolkninger af prøveresultaterne og dels at skaffe den videnskabelige evidens der kan sandsynliggøre at disse antagelser og fortolkninger ikke er helt urimelige (American Educational Research Association, American Psychological Association, National Council of Measurement in Education. 2014). Det ser dog ikke umiddelbart ud til at der eksisterer noget systematisk valideringsprogram for de nationale prøver i Danmark. Uanset hvorfor det forholder sig sådan, virker det uprofessionelt i forhold til internationalt gældende standarder. Jeg ville tro at hvis man systematisk undersøger og løbende offentliggør evidensen for prøvernes validitet, ville man ikke blot kunne løfte prøvernes kvalitet i det lange løb, men også pleje acceptabiliteten af prøverne blandt aftagerne (lærere, elever, forældre etc.) mere effektivt. Det er fx rigtig fint at vi har adgang til at læse om hvordan karakterkriterierne fastsættes. De ser ud til at være testækvivalerede og *ikke normbaserede* (Damvad Analytics, 2016). Det ser ud til at der har været et par kritiske røster mod MCT-prøverne siden de blev indført (Lauritsen, 2006; Andersen & Linderoth, 2012; Lunde & Lunde, 2016), et par kommentarer (Hansen, 2006; Allerup, 2012) og endnu færre eksempler på at de testansvarlige har bragt relevant validitetsevidens til torvs som modsvar på kritik (Nørgaard et al., 2006). Kritikken af muligheden for tilfældigt gætteri blev luftet fra starten hvor Lauritsen (2006) slog på tromme for korrektion for tilfældigt gætteri i MCT. I den forbindelse var det efter min vurdering ret effektivt at Nørgaard et al. (2006) med baggrund i autentiske data og analyser kunne mane den ånd i jorden. Resultaterne af analyser baseret på Item Response Theory (og reproducerbarhedskoefficientens størrelse) understøttede simpelthen ikke antagelser om at tilfældigt gætteri var et stort problem for afgangsprøverne i fysik/kemi i december 2005 (Nørgaard et al., 2006; Downing, 2003a). Og det er ret vigtigt at man husker at *skelne mellem reelle og potentielle problemer* før man springer til at handle ved fx at indføre korrektion for gætteri. Det er jo sådanne overvejelser der er med til at afgøre om man bedriver evidensbaseret eller meningsbaseret pædagogik. Men den slags evidens som Nørgaard et al. (2006) leverede, burde ikke være sporadisk og reaktiv. Den burde selvfølgelig være løbende, systematisk og offentligt tilgængelig for alle afholdte nationale prøver, for nu står vi her igen. Sunde og Sunde (2016) har

også regnet på *potentialet* for at slippe igennem ved tilfældigt gætteri og foreslår igen – alene på den baggrund – korrektion for gætteri. Vi har altså et genfærd, men ingen ghostbusters. Ministeriet for Børn, Undervisning og Ligestilling kan ikke lige henvise til eksisterende tekniske rapporter eller lignende som kan oplyse os om det reelle omfang af tilfældigt gætteri med udgangspunkt i den faktiske evidens fra den afholdte prøve som Sunde og Sundes (2016) indlæg tager udgangspunkt i (personlig kommunikation, 2016).

Lad os i stedet huske på nogle af de reelle udfordringer som eksisterer for vores eksameners validitet. De største trusler mod vores eksameners validitet er "Construct Under-representation", CU, og "Construct Irrelevant Variance", CIV, (Downing, 2002a). Lidt groft formuleret kan man sige at CU referer til situationer hvor testen har et utilstrækkeligt signal, og CIV til situationer hvor der er for meget støj. Det gælder altså om både at fange et tilstrækkeligt signal og samtidig skrue tilstrækkeligt langt ned for støjen hvis transmissionen skal være nogenlunde meningsfyldt.

Det utilstrækkelige signal (CU) er typisk et ret stort problem for de prøveformater som tester meget få indholdsemner i forhold til faget som helhed (Downing, 2002a). Den problematik er typisk et problem for de lange skriftlige prøver (essayformatet) og de mundtlige eksamener når de bliver brugt til at teste viden (Wass et al., 2002). Der er som oftest alt for korte eksaminationstider i disse ellers populære prøveformater til at sikre det der i gamle dage kaldtes indholdsvaliditet, hvis prøverne også skal være gennemførlige i praksis (Wass et al., 2002). Vores elevers og studerendes præstationer er nemlig først og fremmest indholdsspecifikke (content/context specificity), og der findes få/ingen generiske kompetencer, så det er helt og aldeles nødvendigt at stikprøven af de indholdsemner vi tester i vores eksamener, er stor nok til at afspejle fagenes bredde tilstrækkeligt (Eva, 2003; van der Vleuten, 2014). Det er derfor også lidt hårrejsende at erfare at man i visse kredse ser på praktisk-mundtlige prøver som et *automatisk* bedre valg af testformat end MCT (Andersen & Linderoth, 2012; Allerup, 2012). I modsætning til essays og mundtlige eksamener lider MCT nemlig typisk ikke af CU, hvis man altså vælger at udnytte muligheden for en tilstrækkelig dækkende stikprøve af spørgsmål som dette prøveformat muliggør, med ikke alt for lange eksaminationstider. Man bedrager sig selv hvis man tror at MCT kun kan teste faktuel viden. Man kan vælge at teste anvendt viden og højere taksonomiske niveauer i stedet hvis man konstruerer sine MCT's rigtigt (Case & Swanson, 2002).

Tilfældigt gætteri *kan* ganske rigtigt være *en* blandt rigtig mange kilder til CIV eller støj – og dermed *en potentiel* trussel for prøvevaliditeten, specielt i prøver med konstruerede svar som MCT (Downing, 2002a; Kane, 2006). Hvis vi antager at det begreb (construct) man har sat sig for at måle, er biologikundskaber, og det så sidenhen viser sig at i praksis spiller tilfældigt gætteri en stor rolle i besvarelserne, så ville tilfældigt gætteri være et konkurrerende testbegreb (construct) af betydning og en

kilde til støj (CIV). Hvis evidensen derimod viser at tilfældigt gætteri ikke er udbredt i praksis (Nørgaard et al., 2006), må man antage at der ikke eksisterer nogen større trussel for testvaliditeten fra den front. Men hvorfor så ikke korrigere for tilfældigt gætteri for alle tilfældes skyld når vi nu netop ikke har adgang til løbende evidens fra alle tidligere afholdte afgangsprøver i biologi fra de testansvarlige der kan bestyrke os i det generelle fravær af netop denne trussel for prøvevaliditeten? En af de kilder Sunde og Sunde (2016) citerer, giver faktisk en god forklaring på hvorfor *korrektion for tilfældigt gætteri kan ende med at blive en trussel for prøvevaliditeten*. Downing (2003a) påpeger at hvis man indfører korrektion for tilfældigt gætteri, tvinger man eksaminanderne til at forholde sig til risikoen for at svare forkert. Derved bliver prøven pludselig også til et mål for eksaminanders risikovillighed og til deres tiltro/forventninger til egne præstationer eller "self-efficacy" (Bandura, 1997). Med andre ord: Risikovillighed og tiltro til egne evner bliver pludselig til konkurrerende testbegreber i forhold til biologikundskaber. Så længe disse størrelser ikke på forhånd er specificerede som relevante testbegreber for faget af de testansvarlige, er de imidlertid at betragte som CIV, dvs. som kilder til støj og ikke som kilder til signal. Det er også værd at bemærke at Downing (2003a) påpeger med henvisning til anden relevant litteratur at ledende testeksperter derfor er gået bort fra korrektion for tilfældigt gætteri fordi "kuren er mere skadelig end lidelsen". Erfaringerne er at *problemet med tilfældigt gætteri ikke er udbredt i praksis medmindre at prøverne er urimeligt svære, eller at undervisningen har været utilstrækkelig* (Downing, 2003a). Set i det lys peger pilen altså tilbage på de testansvarlige eller underviserne som de potentielle udløsende årsager til støjen – og ikke på eleverne. Det er underviseres opgave at sikre tilstrækkelig kongruens (Constructive Alignment) mellem læringsmål, læringsaktiviteter og eksamen (Biggs & Tang, 2007).

Hvis Sunde og Sunde (2016) virkelig ville kritisere MCT, skulle de ikke have holdt sig til kun at diskutere de matematiske og statistiske aspekter ved tilfældigt gætteri. Andre meget væsentligere kilder til støj eller CIV er meget mere udbredte for dette testformat (Downing, 2002a, 2002b & 2003a). Det er utroligt udfordrende at skrive gode MCT-spørgsmål. Der er virkelig mange muligheder for at lave tekniske fejl der enten tilgodeser testsnuhed eller tilfører irrelevant sværhed til prøven (Case & Swanson, 2002) således at testsnuhed og irrelevant sværhed meget let bliver konkurrerende testbegreber i forhold til eksempelvis biologikundskaber. Forskningen har vist at fejlrater på mellem 28-75 % i MCT-sæt overhovedet ikke er ualmindelige (Downing, 2002b; Downing, 2003b; Palmer & Devitt, 2007; Tarrant & Ware, 2008; Rodrigues-Diez et al., 2016). En anden og beslægtet akilleshæl er at kvalitetssikring af MCT også er meget ressourcekrævende hvis man vil være relativt sikker på at minimere støjen (CIV) fra de tekniske fejl. Vi undersøgte det for nylig og fandt at det krævede mindst 14 erfarne korrekturlæsere for at sikre høje reproducerbarhedskoefficienter (Φ>0,90)

i fejlfindingsprocessen (Mortensen & O'Neill, 2016), og det er ikke ualmindeligt at hver korrekturlæser skal bruge omkring 3 timer på at gennemgå et MCT-sæt med 80 spørgsmål for fejl (Mortensen & O'Neill, 2016; Engelhard et al., 1999). Derfor virkede det umiddelbart betryggende for mig at læse at man i udviklingsprocessen af afgangsprøverne i biologi (i hvert fald tidligere) har ladet hvert MCT-spørgsmål gennemgå mindst 15 bearbejdelsesfaser og ca. lige så mange personer før de bliver godkendt (Nørgaard et al., 2006). Det er sandsynligvis en ret væsentlig årsag til at kun en mindre del af spørgsmålene i sættet (ca. 6 %) ikke opførte sig som forventet til biologieksamenerne i 2005 (Nørgaard et al., 2006). Vi må virkelig håbe at de testansvarliges gode praksis og lærernes gode undervisning er fortsat siden da.

Referencer

American Educational Research Association, American Psychological Association, National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington (D.C.): American Educational Research Association.

Allerup, P. (2012). Folkeskolens centralt stillede test. *MONA, 2012-3,* s. 84-87.

Andersen, P.U. & Linderoth, U.H. (2012). Undervisning og centralt stillede test i folkeskolen. *MONA*, *2012-2*, s. 23-36.

Bandura, A. (1997). *Self-efficacy: the exercise of control*. New York: Freeman.

Case, S.M. & Swanson, D.B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd edition). Philadelphia: National Board of Medical Examiners.

Biggs, J & Tang, C. (2007). *Teaching for Quality Learning at University* (3rd edition). Maidenhead: Open University Press.

Damvad Analytics. (2016). *Forcensur ved folkeskolens 9. klasses afgangsprøver 2015*. Lokaliseret den 3. januar 2017 på: http://www.damvad.com/wpcontent/uploads/2016/05/Endelige-rapport-Forcensur-2015-08.04.2016.pdf.

Downing, S.M. (2002a). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Science Education Theory and Practice*, *7(3)*, s. 235-241.

Downing, S.M. (2002b). Construct-irrelevant variance and flawed test questions: do multiple-choice item writing principles make any difference? *Academic Medicine*, *77(10),* s. 103-4.

Downing, S.M. (2003a). Guessing on selected-response examinations. *Medical Education*, *37(8),* s. 670-671.

Downing, S.M. (2003b).Validity: On meaningful interpretation of assessment data. *Medical Education*, *37(9),* s. 830-837.

Engelhard, G.E., Davis, M. & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, *12(2),* s. 199-210.

Eva, K.W. (2003). On the generality of specificity. *Medical Education*, *37(7),* s. 587-588.

Hansen, P. (2006). Kommentar til "En prøve i bakgear". *MONA*, *2006-1*, s. 94.

Kane, M.T. (2006). Validation. I: R.L. Brennan (red.), *Educational Measurement* (s. 17-64). Westport: ACE/Praeger.

Lauritsen, H.J. (2006). En prøve i bakgear. *MONA*, *2006-1*, s. 1-10.

Messick, S. (1989) Validity. I: H. Wainer & B.H. Braun (red.), *Test Validity* (s. 13-103). New York: American Council on Education og Macmillan.

Mortensen, S.M.R. & O'Neill, L.D. (2016). *Detecting flawed multiple choice items before test administration. A generalizability study*. Manuskript indsendt til publikation.

Nørgaard, K., Steinmüller, L. & Lund-Larsen, M. (2006). De digitale afgangsprøver har høj kvalitet. *MONA*, *2006-3*, s. 86-92.

Palmer, E.J. & Devitt, P.G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? *BMC Medical Education*, *7*, s. 49.

Rodrigues-Diez, M.C., Alegre, M., Diez, N., Arbea, L. & Ferrer, M. (2016). Technical flaws in multiple choice questions in the access exam to medical specialties ("examen MIR") in Spain (2009-2013). *BMC Medical Education*, *16*, s. 47.

Tarrant, M. & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42(2),* s. 198-206.

Van der Vleuten, C.P.M. (2014). When I say … context specificity. *Medical Education*, *48(3)*, s. 234-235.

Wass, V., van der Vleuten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *Lancet*, *357(9260),* s. 945-949.