Sprogevaluering: Begrebets historie og dimensioner

Hvad er sprogevaluering? Hvilke begreber arbejder forskere og sproglærere med for at beskrive og kvalificere sprogevaluering? Det er spørgsmål som denne oversigtsartikel svarer kort på med eksempler fra dansk sprogundervisningspraksis. Formålet med artiklen er at give et historisk og teoretisk overblik over sprogevaluering som fagområde i undervisningssammenhænge hvor elevers læring er i fokus. Evaluering dækker over mange praksisser og tager mange former. Evaluering er nemlig en del af undervisningen, uanset om den har form af observation af deltagelse i en 1. klasses engelsktime eller karakter til folkeskolens afgangsprøver. Evaluering er en helt naturlig del af undervisningen og på samme tid et begreb som kan være svært at afgrænse.

Sprogevaluering bliver et forskningsområde

Selvom evaluering er en helt normal praksis i sprogundervisningen, så er det faktisk et nyere forskningsfelt inden for undervisning (Allerup m.fl. 2011: 38) og mere specifikt sprogdidaktik og pædagogik. Forskningsfeltet opstod i det 19. århundrede i USA (Andreasen m.fl. 2011), hvor forskerne ville måle undervisningsudbytte. Den videnskabelige objektive test blev anvendt i dansk skoleregi fra 1930'erne (Ydesen 2011). Ordet "test" hænger sammen med udviklingen af mental testning eller intelligenstest i starten af det 20. århundrede (Rasborg 1986, bd. 1). I 50-60'erne starter en ny æra i testning, hvor

DEA JESPERSEN
Ph.d.-studerende
Københavns Universitet
deajespersen@hum.ku.dk

psykometri[1] spiller en central rolle (Hamp-Lyons 2016). Det er fx en psykometriker som analyserer cloze-test-prøven i engelsk og tysk til folkeskolens udvidede afgangsprøve (Prien 1981).

I de danske pædagogiske kredse optræder begrebet "evaluering" fra 1960'erne (Borgnakke 2008: 12) fordi forskere manglede et mindre negativt ladet ord til at tale om vurderinger foretaget i en uddannelseskontekst. "Testning" var et ladet ord man gerne ville tage afstand fra (Borgnakke 2008: 17). Begrebet kom fra det amerikanske ord "evaluation" gennem Sverige og Norge (Rasborg 1986, bd. 1: 63). Ordet "evaluering" eksisterede ellers allerede i det danske sprog (Allerup m.fl. 2011: 45), men med en lidt anden betydning. Brugen af ordet tager virkelig til i 1990'erne og frem (Allerup m.fl. 2011: 46-47). I de pædagogiske kredse ses det fra 80-90'erne som et redskab til at udvikle undervisning og skole på et institutionelt niveau (Borgnakke 2008: 18).

Sprogevaluering som forskningsfelt opstod i 1960'erne på den internationale scene med afsæt i 1961 hvor Robert Lado udgav sin bog *Language Testing* (1961) hvis teorier bl.a. blev brugt af en forfatter i Danmark til at analysere realeksamen i engelsk (Høgsbro 1970). En anden stor begivenhed var introduktionen af kommunikativ kompetence bl.a. med Canale og Swains artikel "Theoretical bases of communicative approaches to second language teaching and testing" (1980). Det kan noteres her at både titlen på Lados bog og Canale og Swains artikel indeholder ordet "testing" og ikke "assessment" som er det begreb man møder hyppigst i engelsksproget litteratur nu. Begrebet "assessment" bliver først taget i brug i 1990'erne i den internationale sprogevalueringsforskning og har stort set erstattet "test" eller bliver brugt som et synonym til det. Da den Europæiske Fælles Referenceramme for Sprog fx udkom i 2001, indeholdt titlen ordet "assessment" (*the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR)). CEFR (Council of Europe 2001) og dens Companion Volume (Council of Europe 2018) bliver i dag brugt i mange lande både i og uden for Europa som en standard der knytter nationale test og evalueringer til CEFR så resultaterne har en bredere international brugbarhed, men også til at sætte mål for evaluering i undervisningen. Det skal dog siges at CEFR ikke er en karakterskala, og at det kræver lokal tilpasning af deskriptorerne for at kunne bruge dem i en evalueringskontekst (Dimova m.fl. 2020: 129-130).

Introduktionen af ordet "assessment" markerer en ændring i tilgangen til sprogevaluering: Man går fra test til udvælgelse og kategorisering af elever/studerende til evaluering som redskab til at

fremme læring (Hamp-Lyons 2016). Formålet med at teste er ikke længere kun evaluering *af* læring, men også evaluering *for* læring.

I dansk kontekst kan interessen for evaluering *for* læring og i undervisningen i sprogfagene spores til 1990-2000'erne med publikationer som beskriver evaluering som intern eller løbende. I *Sprogforum 11* om "Evaluering af sprogfærdighed" (1998) er der fx både en artikel om "Løbende evaluering af mundtlig sprogfærdighed i dansk som andetsprog" hos voksne og om "Intern evaluering af elevernes kommunikative kompetence" i engelsk i folkeskolen. Selv i artikler med titler der indeholder ordene "prøver", "test" og "eksamen", som omhandler formelle, eksterne evalueringsformer, er undervisningen, testtagerne og/eller den sociale kontekst med i overvejelserne. Det skal dog her understreges at den analyse er lavet på et smalt grundlag da der kun eksisterer få danske publikationer om sprogevaluering.

Afrundende kan man sige at tendenserne inden for sprogevaluering følger tendenserne inden for evaluering (Dahler-Larsen 2011: 11) og inden for sprogundervisning (Green 2013: 174). I starten kiggede man mest på selve testens interne virkning, dvs. om den testede det den skulle, og senere hen er man begyndt at se mere bredt på testens effekt i undervisningen, på læring og i samfundet. Evalueringsformerne udviklede sig fra at være primært baseret på *discrete point items* testet i multiple choice-formatet til at inkludere mange andre former, fra rollespil til portfolio med fokus på alt fra viden til kompetencer.

At definere evaluering

Da meget af litteraturen inden for evaluering kommer til Danmark gennem engelsksproget litteratur, er mange af begreberne man bruger til at tale om evalueringsformer, enten låneord eller oversat til dansk. Men det betyder ikke at deres betydning er helt ens, eller at der er enighed om deres betydning (Allerup m.fl. 2011: 41). Det er et dynamisk begreb i den forstand at det er under konstant udvikling og skal omfavne mange former for praksisser.

Betydningen af begrebet er tæt beslægtet med vurdering og testning. Carlsen og Moe (2019) mener at det norske ord "vurdering" svarer til begrebet "assessment" som bliver brugt inden for international forskning: "*Vurdering (assessment)* er det overordnede begrepet og omfatter all form for innsamling av språklig data eller observasjon av språkbruk"[2] (Carlsen & Moe 2019: 28). Den definition indfanger således bredt alle de former for vurderinger, som finder sted i undervisningen. Det kan være alt fra at læreren ser elever i 1. klasse reagere

på instrukser på engelsk, til retning af stile i 3.g. Allerup m.fl. (2011) mener dog at vurdering på dansk er en delproces af evaluering, altså at "evaluering" er det overordnede begreb. Evaluering sker når der indsamles data systematisk, og der bliver draget konklusioner "i form af en karakter eller en udtalelse ud fra et mere eller mindre veldefineret værdigrundlag, hvor der bruges et eller flere måleredskaber. Disse måleredskaber kan være knyttet til en værdiskala, fx trinmål, karakterskala, e.a." (2011: 162). Man sætter en værdi på fx elevens sprogbrug. I den hensigt kan man sige at Allerup m.fl.'s forståelse af evaluering nærmer sig Brown og Abeywickramas definition af det amerikanske begreb "evaluation" (Brown & Abeywickrama 2019: 5). De ser *evaluation* som det at give en værdi (*value*) til målinger. De tilføjer at *evaluation* sker når resultaterne[3] afgør om en elev fx er dumpet eller ej, eller om det er et godt stykke arbejde eleven har lavet. Sammenfattende kan man sige at evaluering kan defineres som "the process of systematically gathering data from learners to make interpretations about their language abilities and decisions about their future" (Chapelle & Voss 2016: 116), hvilket ironisk nok er en definition givet på ordene "language assessment" og "language testing", men som i denne sammenhæng passer på evaluering.

"Testning" er det andet ord som er tæt forbundet med evaluering. Der er mere konsensus i litteraturen om hvad test er for en størrelse. Det er en evalueringsform (eller vurderingsform) som "gjør bruk av et måleinstrument (en test) som er utviklet i henhold til klare regler og prosedyrer, og som skiller seg fra andre mer uformelle former for vurdering blant annet ved at man har en klar plan for hvilke språkferdigheter eller språklige komponenter man ønsker å observere og kvantifisere" (Carlsen & Moe 2019: 30). En test er således en af de mange former evaluering kan tage.

"Test" er faktisk en relativt ny term som hænger sammen med udviklingen af mental testning eller intelligenstest i starten af det 20. århundrede (Hamp-Lyons 2016; Rasborg 1986, bd. 1). Eksamener og prøver, derimod, kan spores helt tilbage til Kina 2000 f.Kr. (Rasborg 1986, bd. 1). I Danmark blev årsprøver brugt fra folkeskolens start i 1814 til 1958 til at bestemme om en elev skulle rykke op til næste årgang (Andreasen m.fl. 2015: 11). Hyppig brug af eksamener og prøver i uddannelsessystemet er ikke noget specielt nyt, jf. udtalelse fra en skolemand i 1939: "Der findes jo intet andet Land, hvor Eksamensglæden er saa stor som i Danmark, vi er helt tossede i vores Tyrkertro paa Værdien af den Slags" (i Andreasen m.fl. 2015: 11).

Men uanset om der er tale om test, prøver, eksamener eller daglig

observation af ens elevers handlinger i undervisningen, så er der tale om evaluering.

At beskrive evaluering

Med tiden er ordforrådet til at beskrive evaluering blevet mere nuanceret, hvilket dels afspejler udbredelsen af evaluering af sprog som praksis, men også en forståelse for fænomenets kompleksitet (Hamp-Lyons 2016: 13). Evaluering kan finde sted på mange niveauer og tage mange former. For at sætte ord på kompleksiteten ved fænomenet er der blevet tilføjet kategorier for at kunne skelne mellem forskellige typer evaluering. Herunder følger nogle af de adjektiver der bliver brugt til at beskrive evaluering.

En vigtig forskel findes i hensigten med evaluering. Skal evalueringen støtte elevers læring, dvs. forme deres læring, så taler man om formativ evaluering. Skal evalueringen derimod informere om en elevs færdigheder, viden og/eller kompetencer i slutningen af et forløb, dvs. lave summen af den læring der har fundet sted, så taler man om summativ evaluering (se også Stæhr 2011: 20). Prøver og eksamener er typisk summative. Men selvom om man kan skelne mellem de to typer af evaluering, så er skellet heller ikke så skarpt endda, da der nærmere er tale om et kontinuum. En lærer kan fx godt bruge en afgangsprøve formativt ved fx at få eleverne til at tale efterfølgende om hvordan de har løst lytteopgaven, og hvilke strategier der kan være de mest hensigtsmæssige. På samme måde kan en opgavetype som man ville tro var formativ som portfolioen, blive brugt summativt hvis den ikke bruges undervejs til at hjælpe med en progression, men blot er til for at dokumentere den.

Adjektiverne "summativ" og "formativ", som kommer fra amerikansk litteratur, blev i 1970'erne afløst i Danmark af "ekstern" og "intern" evaluering (Rasborg 1986, bd. 1: 68-75). Termerne har siden da fået deres eget liv parallelt med summativ og formativ evaluering som også bliver brugt til at beskrive evaluering. I ekstern evaluering bliver resultaterne brugt af udefrakommende. Eksamener og prøver er ekstern evaluering. Intern evaluering sker i undervisningen eller bliver brugt af den instans der har brug for evalueringen. For eksempel er en mundtlig præsentation i klassen er intern evaluering. Men her igen kan det være svært at lave et hårdt skel mellem de to former for evaluering: Terminsprøver er generelt en eksternt udviklet prøve, hvor karakteren bliver brugt internt.

For at tale om hvordan evaluering påvirker interessenter (fx ele-

ver, lærere, forældre, skoleledere m.fl.), har danske forskere taget udtrykkene "high stakes" og "low stakes" til sig. Evaluering er "high stakes" når der er meget på spil fx i form af adgang til statsborgerskab eller uddannelse. Da karaktererne til de afsluttende prøver i folkeskolen blev adgangsgivende til en ungdomsuddannelse i 2017, kom der lige pludselig mere på spil for 9. klasse-eleverne. Folkeskolens prøver blev "high stakes". Er der tale om samtaler i klassen, hvor læreren observerer eleverne, er der sandsynligvis mindre på spil for alle parter. Evalueringen er "low stakes".

Graden af ensformighed af proceduren kan også beskrives. Mange af de prøver og test som er "high stakes", eksterne og summative, er standardiserede, dvs. at der er ensartethed i testadministration, alle testtagere får samme oplæg, og der bliver anvendt faste procedurer samt karakterskalaer. Størstedelen af evalueringen der sker i klasserummet, er ikke standardiseret.

Mens disse termer hjælper os med at skelne mellem evalueringer, så giver de os ikke en ide om kvaliteten af en evaluering.

At kvalificere evaluering

Et centralt begreb inden for sprogevaluering er "validitet" (Luoma 2002), som nogle gange kaldes "gyldighed" på dansk. Det er også et omdiskuteret koncept. Validitet handler om i hvor høj grad resultaterne af en evaluering er retvisende i forhold til dens formål, og i hvor høj grad brugen af resultaterne er hensigtsmæssig (Messick 1989). Imidlertid er problemet med det begreb at det blev udviklet til at tale om kvaliteten af evalueringsformer som generelt var "high stakes", standardiserede og summative (Moss m.fl. 2006: 112). Det er det meste af klasserumsevaluering ikke. Bachman og Palmer (1996) og Brown og Abeywickrama (2019) har konkretiseret begrebet ved at omsætte det til kvalitetskriterier inden for sprogevaluering for at vurdere evalueringers brugbarhed (se fx Rønsholt & Pinholt 1998 for en analyse af en prøve ud fra de principper).

Et af kriterierne er meget praktisk og handler om gennemførlighed. Kan evalueringen gennemføres under de praktiske forhold? En elevproduceret film med detaljeret feedback på form og indhold kan være meget lærerig, men måske tage alt for lang tid at give respons på. I så fald er gennemførligheden udfordret.

En anden kvalitet er indvirkningen som gerne skulle være positiv. Hvordan påvirker evalueringen eleverne eller mere generelt undervisningen? Hvis en evalueringsform har en negativ effekt på fx elevers læring eller på undervisningens planlægning, så taler man om

negativ "washback". Virker en evalueringsform derimod motiverende på eleverne, og hænger den godt sammen med undervisningen, er der tale om mulig positiv "washback". Problematikken omkring "washback" (eller "backwash") har været meget omdiskuteret i forbindelse med de nationale test på folkeskoleniveau (se fx Stæhr 2011).

Særligt vigtigt i kommunikativ sprogevaluering er autenticitetsprincippet som handler om at sproget der bliver evalueret, er så tæt på reel sprogbrug som muligt. Det princip kan være særligt svært at opnå i grunduddannelser, da det ikke altid lige er til at vide hvad eleverne kommer til at møde af sprogbehov uden for skolen.

Et princip som traditionelt har fyldt mest når man taler om test, er reliabilitet, også kaldt pålidelighed. Er en test pålidelig, vil to elever på samme sproglige niveau og under samme testforhold få det samme resultat. Testen i sig selv skal være pålidelig, dvs. at den skal være ordentligt designet og foretages under forhold som er så ens som muligt. Hvis eleverne bagerst i klassen ikke kan høre lige så godt som eleverne på forreste række i en lytteøvelse, så er evalueringens pålidelighed udfordret. Men reliabiliteten kan også være påvirket af faktorer som nervøsitet hos eleven eller træthed hos læreren. Måske ville eleven klare sig bedre hvis han/hun var mindre nervøs, eller måske ville en skriftlig opgave have fået en lavere karakter hvis den havde ligget øverst i bunken.

Konstruktvaliditet eller begrebsvaliditet sikrer at man måler det man har sat sig for at måle. Det vil sige at hvis man ønsker at få indsigt i sine elevers skriftlige færdigheder, så vil en multiple choice-grammatikprøve ikke række. Konstruktet er den teoretiske definition man har af den færdighed eller kompetence man ønsker at evaluere. Måler evalueringen det den skal? Et eksempel af konstruktvaliditet der er gået galt, kunne være at man ønsker at se på 4. klassers skrivefærdigheder i engelsk ved at få dem til at skrive en lille tekst om deres familie på computeren. 4. klasserne er imidlertid ikke vant til at bruge computer, og meget energi bliver brugt på digitale færdigheder frem for skrivning på engelsk.

Meget forskning inden for sprogevaluering interesserer sig for validering, dvs. at kunne argumentere for at resultaterne fra en test faktisk kan bruges hensigtsmæssigt (Luoma 2002). Det skal understreges at selv efter en grundig valideringsproces vil der altid vil være områder der kan forbedres i sprogevaluering. Den perfekte test eksisterer ikke. Og evaluering kan ikke tages ud af den kontekst den bliver givet i.

Evaluering på godt og ondt

Afsnittets titel er også titlen på en formidlingsbog af J. Dolin (2020) om evaluering, i hvilken han tager udgangspunkt i de gavnlige og mindre hensigtsmæssige aspekter af evaluering. Evaluering er en del af praksis i alle uddannelseskontekster og har vist sig at være et vigtigt element af læring bl.a. i form af feedback. Det er derudover et fænomen som er svært at komme uden om fordi det er socialt forankret. Men evaluering har også en mindre positiv side. De seneste år er der kommet mere og mere fokus på karakterer, test og prøver, hvilket kan påvirke undervisning og læring negativt bl.a. i form af *teaching to the test* og stressede elever. Evaluering kan være et magtfuldt politisk styrings- og kontrolredskab.

Sprogevaluering handler også om magt. Resultaterne af en sprogprøve kan være nøglen til permanent opholdstilladelse i Danmark eller dansk statsborgerskab. I klasserummet kan elever opleve evaluering som disciplinerende og kontrollerende (Shohamy 2002: 32), hvilket kan sætte spor. Den magt og den indvirkning evaluering har på mennesker og samfundet, bliver taget mere og mere alvorligt. Det har udviklet sig til den internationale bevægelse *critical language testing* eller kritisk sprogtestning (Shohamy 2002: 34) som startede i 1990'erne med Shohamys arbejde om sammenhæng mellem sprogtest, magt og demokrati. Kritisk sprogtestning handler om at sætte spørgsmålstegn ved måden evaluering bliver brugt på, og de konsekvenser den har for mennesker og samfund. Det har bl.a. ført til udvikling af etiske kodekser blandt internationale sprogevalueringsforeninger som EALTA, ALTE og ILTA (se Saville & Pedersen 2002).

Lærere/undervisere og elever/studerende/kursister er centrale i på ene side at sikre en mere hensigtsmæssig og kvalificeret brug af evaluering i sprogfagene, og på den anden side at støtte evaluering der fremmer læring. Det har givet anledning til en ny, vigtig praksis der har fundet sted inden for sprogevaluering siden 2010'erne, nemlig at udvikle læreres *language assessment literacy* eller litteracitet i sprogtestning og evaluering på dansk (se artikler i dette nummer af *Sprogforum*). I Danmark vil sproglærere gerne evaluere, dvs. indsamle systematisk viden om deres elevers læring, med det formål at støtte elevernes læring. Men hvor lærerne ikke mangler viljen, mangler de redskaberne (Danmarks Evalueringsinstitut [EVA] 2003; Poulsen 1998).

Afrunding

Dette er fjerde gang at tidsskriftet *Sprogforum* har evaluering som tema. Karen Lund skrev i introduktionsartiklen til *Sprogforum* "Test!" (2011) at der fortsat er stort fokus på summativ evaluering når det kommer til forskning og omtale. Selvom prøver og test stadig fylder, er der i dag mere og mere fokus på evaluering der fremmer læring og er mere kontekstbaseret. Det understreger hvor vigtigt det er også at fokusere på litteracitet i sprogtestning og evaluering både for at kunne forholde sig kritisk over for ekstern evaluering og for at kvalificere intern evaluering.

Slutnoter

1 Psykometri beskæftiger sig med kvantificerbar måling af psykologiske og kognitive træk.
2 Kursiv er fra originalteksten.
3 Resultater kan være kvantitative (fx i form af karakter) eller kvalitative (fx i form af udtalelse).

Litteratur

Allerup, P., Jansen, M. & Weng, P. (2011). *Evaluering i skolen, baggrund, praksis, teori*. Frederikshavn: Dafolo.

Andreasen, K., Buchardt, M., Rasmussen, A. & Ydesen, C. (2015). Test som historisk og socialt fænomen. I: Andreasen, K., Buchardt, M., Rasmussen, A. & Ydesen, C. (red.), *Test og prøvelser, oprindelse, udvikling, aktualitet* (s. 9-25). Aalborg: Aalborg Universitetsforlag.

Andreasen, K., Rasmussen, A. & Friche, N. (2011). Evaluering i uddannelsespolitik og praksis. I: Andreasen, K., Friche, N. & Rasmussen, A. (red.), *Målt & vejet: Uddannelsesforskning om evaluering* (s. 9-35). Aalborg: Aalborg Universitetsforlag. forskning.ruc.dk/da/publications/evaluering-i-uddannelsespolitik-og-praksis.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Borgnakke, K. (2008). Evalueringsstrategier i den pædagogiske kontekst. I: Borgnakke, K. (red.), *Evalueringens spændingsfelter* (s. 9-66). Aarhus: Klim.

Brown, H.D. & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3. udg.). Hoboken: Pearson.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1), 1-47. doi.org/10.1093/applin/I.1.1.

Carlsen, C. & Moe, E. (2019). *Vurdering av språkferdigheter*. Bergen: Fagbokforlaget.

Chapelle, C.A. & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology*, 20 (2), 116-128. https://doi.org/10125/44464.

Council of Europe (2001). *Common European framework of reference*

for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press. Lokaliseret d. 18. August 2019 på coe.int/en/web/common-european-framework-reference-languages/home.

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment; companion volume with new descriptors*. Strasbourg: Council of Europe. Lokaliseret d. 18. august 2019 på coe.int/en/web/common-european-framework-reference-languages.

Dahler-Larsen, P. (2011). Sprogtest, evalueringer og deres virkninger. *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 17 (52), 11-18. tidsskrift.dk/spr/article/view/102824.

Danmarks Evalueringsinstitut [EVA] (2003). *Engelsk i grundskolen – om indhold, organisering og vilkår*. Lokaliseret d. 5. juni 2020 på eva.dk/grundskole/engelsk-grundskolen-om-indhold-organisering-vilkaar.

Dimova, S., Ginther, A. & Yan, X. (2020). *Local language testing: Design, implementation, and development*. New York: Routledge.

Dolin, J. (2020). *Evaluering på godt og ondt*. Aarhus: Aarhus Universitetsforlag.

Green, A. (2013). Multiplication and division – trends in language assessment. I: Green, A., *Exploring language assessment and testing, language in action* (s. 171-220). London: Routledge.

Hamp-Lyons, L. (2016). Purposes of assessment. I: Tsagari, D. & Banerjee, J. (red.), *Handbook of second language assessment* (bd. 12, s. 13-27). Boston: De Gruyter Mouton. doi.org/10.1515/9781614513827-004.

Høgsbro, A. (1970). *Realeksamen i engelsk*. København: Gyldendal.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. London: Longmans, Green and Co.

Lund, K. (2011). Forord – test! *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 17 (52). tidsskrift.dk/spr/article/view/102822.

Luoma, S. E. (2002). Hvad sker der inden for sprogtestning? (M. S. Pedersen, Overs.). *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 8 (23), 7-15. tidsskrift.dk/spr/article/view/113295.

Messick, S. (1989). Validity. I: Linn, R. L. (red.), *Educational measurement* (3. udg., s. 13-103). Washington, DC: American Council on Education/Macmillan.

Moss, P. A., Girard, B. J. & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 30 (1), 109-162. doi.org/10.3102/0091732X030001109.

Poulsen, E. (1998). Intern evaluering af elevernes kommunikative færdigheder: – Et forskningsbaseret udviklingsarbejde i engelsk i folkeskolen. *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 4 (11), 33-38. tidsskrift.dk/spr/article/view/116654.

Prien, B. (1981). *Cloze-tests i tysk og engelsk: En undersøgelse af to prøver fra folkeskolens udvidede afgangsprøve 1979*. København: Danmarks Pædagogiske Institut.

Rasborg, F. (1986). *Intern evaluering* (bd. 1-3). København: Danmarks Pædagogiske Institut.

Rønsholt, S. & Pinholt, P. (1998). Løbende evaluering af mundtlig sprogfærdighed i dansk som andetsprog. *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 4 (11), 19-25. tidsskrift.dk/spr/article/view/116650.

Saville, N. & Pedersen, M. S. (2002). Kvalitet og retfærdighed:

ALTE's Code of Practice. *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 8 (23), 45-50. tidsskrift.dk/spr/article/view/113301.

Shohamy, E. (2002). Demokratiske og udemokratiske dimensioner i evaluering. *Sprogforum*, 8 (23), 31-36.

Stæhr, L.S. (2011). De nationale test – positiv eller negativ washback på undervisningen? *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 17 (52). tidsskrift.dk/spr/article/view/102825.

Ydesen, C. (2011). Risikofyldte test – erfaringer fra dansk testhistorie. *Sprogforum. Tidsskrift for sprog- og kulturpædagogik*, 17 (52), 41-45. tidsskrift.dk/spr/article/view/102828.