Hvad sker der inden for sprogtestning?

Sari E. Luoma

Forsker på Jyväskylä Universitet i Finland og evalueringsspecialist ved Defense Language Institute Foreign Language Center (DLIFLC).

Sprogundervisningens og sprogtestningens verdner har ændret sig radikalt inden for de sidste 20-30 år. Med den kommunikative revolution blev sprogtilegneres erfaringer med at lære sprog i institutionaliserede sammenhænge eksplicit forbundet med *brugen af sproget*, og lige siden er der blevet introduceret nye måder til at hjælpe sprogtilegnere med at opnå kommunikative færdigheder. Vi har set hvordan der bl.a. er blevet lagt vægt på dialoger og pararbejde, anvendelse af sproget i bestemte kontekster, strategi-træning, og udvikling af interkulturel bevidsthed.

Med en vis forsinkelse og i modificeret form har de samme træk fundet vej til praksis inden for testning og evaluering. Forsinkelsen kan delvis forklares med den iboende konservatisme der ligger i testning som en normbaseret aktivitet. Men den skyldes også til en vis grad lærer-bedømmerens afgørende rolle i planlægningen og gennemførelsen af evaluering, kombineret med en bevidst eller ubevidst frygt for de strenge krav til korrekthed og retfærdighed som ligger i evaluering. Det er nemmere at følge en gammel praksis og tro på at den har stået sin prøve, end det er at se sin frygt og usikkerhed om evaluering i øjnene. Denne situation er imidlertid ved at ændre sig.

Mens hovedstrømmen stadig lægger vægt på måling og standarder, er der inden for det seneste årti sket ændringer i evalueringsteorier og -praksisser som berører alle områder inden for metode, teori og aktiviteter (en tilgængelig oversigt findes f.eks. i Birenbaum 1996), og det bliver stadig tydeligere at der i dag er flere forskellige strømninger

inden for sprogtestningen. Fra at have været begrænset til *måling* har sprogtestning udvidet sig til *sprogevaluering*, et område med funktionelle tilknytninger til sproguddannelse, en række områder inden for anvendt lingvistik og uddannelsesevaluering.

I denne artikel vil jeg diskutere nogle af de ændringer der har fundet sted inden for sprogtestning og -evaluering i den 'traditionelle' betydning og i den nyere, bredere betydning. Artiklen fokuserer på problemstillinger og implikationer der er relevante for sproguddannelse, men jeg vil også omtale nye artikler og oversigter som giver en mere detaljeret forklaring af særlige udviklinger.

Teknologisk udvikling: Computerbaseret fleksibel testning

Computere og internet udgør en vigtig kontekst for sprogtilegneres brug af sprog i dag. Dette gælder især for sprogundervisning i engelsk, skønt anvendelsesområdet, især med hensyn til specialiseret sprogindlæring, også er blevet udviklet til andre sprog. De spændende muligheder for sprogevaluering som anvendelsen af computere giver, ligger i udviklingen af computerbaserede fleksible test- og evalueringsinstrumenter.

Computerbaserede fleksible evalueringsværktøjer indeholder for-programmerede opgavesæt som kan organiseres efter emne, færdighed der skal testes, og / eller særlige emneområder som skematiske spørgsmål-svar-sekvenser. Derudover gives der oplysninger om måling i forbindelse med opgaverne. Disse oplysninger angiver hyppigst hvor vanskeligt et item har vist sig at være i præ-tests og analyser.

Idéen med computer-baserede fleksible tests er at opgavegivningen bliver tilpasset den enkelte sprogtilegners svar. Tilegneren får først en opgave - også kaldet et *item* - på mellemniveau. Hvis han eller hun har løst opgaven rigtig, giver programmet en sværere opgave. Hvis også denne besvarelse er korrekt, vælger programmet et endnu sværere item. Hvis tilegneren nu laver fejl, giver programmet et lettere item og begynder at indkredse tilegnerens færdighedsniveau. En anden tilegner kan få den samme opgave til at begynde med, men kan - afhængig af sit besvarelsesmønster - få helt andre sæt af items. Programmet kan ikke desto mindre bruge de målingsoplysninger der er knyttet til de enkelte items, til at give en endelig karakter til begge tilegnere på den samme skala.

Det statistiske grundlag for at lave computer-baserede fleksible tests er for det meste en eller anden form for *Item Responsteori* (IRT). IRT-modeller er baseret på sandsynligheden for korrekt svar, med andre ord sandsynlighedsteori[1].

Fordelene ved computer-baseret fleksibel testning er bl.a. at tests kan være kortere når den ekstra tid der går med at besvare items som er for nemme eller for vanskelige for sprogtilegneren, spares, og at resultater fra tests der er blevet givet på forskellige tidspunkter, er sammenlignelige. Når først et lager af items er blevet afprøvet og analyseret således at item-værdierne er kendte,

kan enhver kombination af dem gives til nye test-tagere. Test-sikkerheden bliver også forbedret når indholdet i en hvilken som helst ny test er uforudsigelig.

Der er imidlertid også nogle ulemper. Den vigtigste praktiske ulempe er omfanget af det arbejde der kræves for at oprette et item-lager til en fleksibel test. Det antal af sprogtilegnere der skal bruges til at afprøve et item-lager ordentligt, kan tælles i hundreder, eller for nogle statistiske modellers vedkommende i tusinder. Udviklingen af fleksible tests er med andre ord ikke noget en hvilken som helst lærer umiddelbart kan gå i gang med.

Hvis brugbare fleksible evalueringssystemer imidlertid allerede eksisterer, kan lærere afprøve dem på deres sprogtilegnere og evaluere deres brugbarhed. Tilegnerne vil sandsynligvis finde det interessant at finde ud af hvordan de bliver indplaceret af forskellige programmer, men man kan også bede dem reflektere over andre aspekter af evalueringsprocessen, som f.eks. hvor sjov den har været, og hvor nyttig den har været som en støtte til deres læring.[2]

Foruden *Item Responsteori* og dens anvendelse i fleksibel testning har der også været andre tekniske og statistiske udviklinger inden for sprogtestning i de seneste 30 år. For dem som er interesseret i disse aspekter af sprogevaluering, findes korte og præcise opsummeringer af udviklingerne i Bachman & Eignor 1997 og Bachman 2000. Ud over information giver disse artikler også yderligere henvisninger til specifikke statistiske teknikker.

Teoretiske udviklinger: Fremgang for validering

Inden for sprogtestningsteori er de vigtigste udviklinger i de seneste årtier sket inden for validering. Validitet har altid været et af de centrale kriterier for kvalitetsmåling i testning, men i første halvdel af det 20. århundrede blev det normalt opfattet som et uproblematisk begreb. Spørgsmålet var om testen udfyldte sin opgave, og dette spørgsmål blev besvaret ved at sammenholde testresultaterne med evaluering af test-tagerens præstation i forbindelse med den opgave som testen rettede sig mod i den virkelige verden. Nutidens validitets-begreb er imidlertid meget bredere. Det drejer sig om hvorvidt testen kan forsvares videnskabeligt og socialt. Dette lyder flot, men hvad test-udviklerens arbejde angår, drejer det sig faktisk om nogle ret praktiske spørgsmål. Validitet har med test-resultaternes betydning at gøre, og det centrale spørgsmål er: „Hvad er test-måling?" efterfulgt af: „Hvordan kan man vide det?".

Valideringsarbejdet begynder med testudviklerens definition af den færdighed som skal evalueres i testen. Rent operationelt bliver færdigheden indarbejdet i opgaverne, evalueringskriterierne og de faktiske aktiviteter der udføres under evalueringen. Disse aktiviteter må nødvendigvis undersøges med henblik på at tjekke at den færdighed man regner med bliver evalueret i testen, også rent faktisk er den som testen retter sig mod. Men i nutidens teorier bliver det understreget at udviklerne må gå længere end det, at de må være i stand til at sige hvad resultaterne betyder.

Skønt dette er et ganske udfordrende krav, er test-udviklerne i en ideel position hvad angår beskrivelsen af de færdigheder de evaluerer, fordi de under udviklingen arbejder med test-opgaverne og evalueringskriterierne i lang tid. I dette arbejde er udviklerne bevidst opmærksomme på *hvordan* den færdighed *det er hensigten* at evaluere, bedst kan operationaliseres inden for evalueringssituationens praktiske begrænsninger. 'Det eneste' de behøver at gøre for at dette arbejde kan blive til en del af valideringen, er at nedskrive deres idéer og grundene til deres beslutninger. Arbejdet fortsætter i form af empiriske undersøgelser af om opgaverne og evalueringskriterierne faktisk indarbejder disse intentioner. Validering indebærer med andre ord mange forskellige metodologier, der alle har som mål at vise hvad testen tester.

Da valideringen imidlertid ikke alene fokuserer på hvad testen tester, men også på betydningen af scoren, bliver de sociale dimensioner i evaluering og score også relevante. Fra en filosofisk synsvinkel indebærer dette en erkendelse af at evaluering er en menneskelig aktivitet der udføres på grundlag af velovervejede vurderinger. Fra en samfundsmæssig synsvinkel indebærer det en erkendelse af at test-scoring ofte udgør et grundlag for samfundsmæssige beslutninger. Når scoren bliver brugt - til at give karakter eller til at beslutte hvorvidt en person er kvalificeret til at studere på et universitet f.eks. - har dette konsekvenser for test-tagerne, og evalueringsudviklere er ifølge den aktuelle validitetsteori delvis ansvarlige for disse konsekvenser. De deler ganske vist ansvaret med dem der bruger scoren, f.eks. de universitetskommissioner der udvælger studerende, men de kan ikke fuldstændig undgå at have et ansvar.

Valideringsarbejde er vigtigt fordi det handler om kvalitet og om ansvarlighed. Hvis selvrefleksionsaspektet af validering tages alvorligt, hjælper dette testudviklerne med at forbedre deres tests. I sin nuværende komplekse form står valideringsteori imidlertid over for en interessant udfordring. Det er nemlig ikke klart hvem det er der har brug for de komplekse valideringsrapporter og -studier som den nuværende teori synes at stille krav om. Hvem andre end en anden tester vil være i stand til at læse rapporterne og være i stand til at vurdere om testen er omhyggeligt udarbejdet og tester det den siger den tester? Den nuværende tendens inden for valideringsteori er derfor at gøre konceptet klart og operationaliserbart[3].

Når validering bliver opfattet som en proces - hvor man bevæger sig fremad ét skridt ad gangen - fører den til ærlige forsøg på at gøre et godt stykke bedømmelsesarbejde - med andre ord give en ansvarlig (*accountable*) bedømmelse. Argumentet for at fortsætte med forskellige former for valideringsundersøgelser er at kvalitet er vigtig. Desuden ville afskaffelse af tests ikke afskaffe behovet for at træffe sociale beslutninger. De ville bare blive truffet på et andet grundlag. Målet med ansvarlig bedømmelse er at fremskaffe et principielt grundlag for kvalificerede beslutninger. Når beslutninger træffes på grundlag af scorerne i sprogtests, er det normalt dem der har de bedste scorer der får

gavn af det. Her er det at validering betyder noget. Det er vigtigt at vide hvad forskellen er på dem med en lavere score og dem med en højere score, således at det kan blive tjekket at beslutningerne er truffet på et forsvarligt grundlag.

Ændringer i omfang og filosofi: Fra testning til bedømmelse

Tidligere havde man den opfattelse at det først og fremmest drejede sig om tests i stor skala med særlig vægt på statistisk analyse. Mens det stadigvæk er ét af de vigtigste aktivitets- og forskningsområder, er et andet lige så vigtigt område læringsrelateret bedømmelse. De bedømmelsesformater der her er relevante, omfatter forskellige former for selvevaluering og partnerevaluering (*peer evaluation*), portfoliobedømmelse, læringsdagbøger osv. Hvad angår testteori, har udviklingen betydet at sprogtestere har måttet revurdere deres antagelser om dette at bedømme, såvel som deres antagelser om de kvalitetskriterier der gælder i forskellige sammenhænge.

Som aktivitet er læringsrelateret bedømmelse mere kompleks og fleksibel end traditionel testning. Afhængig af behovene i de forskellige bedømmelsessituationer kan organiseringen variere med hensyn til:

- *Hvem bedømmer?* Læreren, andre studerende, sprogtilegnerne selv?
- *Hvad bedømmes?* Det enkelte svar på et spørgsmål eller en opgave? Den generelle præstation i klasselokalet? Indstilling til sprogtilegnelse og sprogbrug? Erfaringer med fremmedsprog uden for klasseværelset?
- *Hvem udvikler opgaven?* Læreren alene? Læreren og de studerende i fællesskab? De studerende i fællesskab? Én studerende til en anden? Den enkelte studerende til sig selv?
- *Hvem træffer beslutning om kriterierne for præstationen?* Læreren alene? Læreren og de studerende i fællesskab? De studerende i fællesskab?
- *Hvad sker der med bedømmelsesresultaterne?* Bliver de brugt som information til læreren og de studerende? Bliver de anvendt som støtte til sprogtilegnernes bevidsthed om sprogets natur og deres brug af sproget? Bliver de anvendt som grundlag for karaktergivning? Bliver de anvendt til vurdering af kursusplanens hensigtsmæssighed? Andre ting?

Den type af kvalitetskriterier der er drivkraften i de valg der træffes inden for læringsrelateret bedømmelse, omfatter ifølge McNamara (2001) at undervisningsprocessen er vedkommende, at læringen bliver faciliteret på en lang række forskellige måder, at kvaliteten af undervisningen forøges, at lærernes administrative byrder nedsættes. Stillet op på denne direkte måde er disse kriterier selvindlysende for lærerne, men de står helt tydeligt i modsætning til de bedømmelseskriterier som testere og uddannelsesadministratorer, såsom undervisningsministerier, har opstillet.

For testteoretikere er det vigtigste kvalitetskriterium det aktuelle brede validitetskoncept. Det betyder at de færdigheder der bliver bedømt, bør defineres, og at definitionen bør kunne forsvares intellektuelt; at der bør være

bevis på reliabilitet; at der også bør være andre former for dokumentation og data der støtter testens validitet - såsom bevis på kvalitetskontrol fra testsudviklingsprocessen og analyser af de studerendes præstationer - for at vise at opgaven faktisk tester de færdigheder som det var hensigten at teste. Desuden bør konsekvenserne af at anvende testen overvejes før proceduren bliver anvendt og bagefter kontrolleret. For administratorer giver bedømmelse og evaluering nogle mål for ansvarlighed. De sætter ministerier i stand til at se hvor store fremskridt der gøres i læring og undervisning, og dermed hvad regeringer og stater får ud af deres investeringer i uddannelse. McNamara påpeger at de tre gruppers behov delvis er i konflikt med hinanden.

Tidligere har praktikere og teoretikere undgået konflikten ved ikke at betragte læringsrelateret bedømmelse som en del af 'test-' og 'administrations'-verdenen. De to grupper har - bevidst eller ubevidst - haft en tendens til at støtte hinanden. Når læringsrelateret bedømmelse nu er ved at blive en del af test-/bedømmelsesverdenen, er der brug for at sikre at foreningen sker i begge parters interesse. Dette indebærer både udfordringer og løfter. Udfordringerne på test-siden omfatter teoretikernes behov for at finde ud af hvilken forskning der er nødvendig for at forstå principperne i læringsrelateret bedømmelse og støtte dens berettigelse. McNamara (2001) henleder f.eks. opmærksomheden på at vi ved meget lidt om de refleksions- og analyseprocesser lærere og studerende gennemgår når de bedømmer en students arbejde, og at processerne kan adskille sig fra hinanden i forskellige bedømmelsessituationer. Udfordringerne på undervisningssiden består bl.a. i at der er behov for eksplicit refleksion over karakteren af de studerendes færdigheder på forskellige færdighedsniveauer, og over de træk ved præstationen (og muligvis bedømmelsessituationen) der fører til forskellige bedømmelsesresultater. Den nye tilgang inden for forskning, refleksion og analyse giver imidlertid løfte om at der vil blive produceret ny viden om de principper der ligger bag vores daglige handlen. Dette vil forhåbentlig forbedre vores arbejde og muligvis også mindske kløften mellem teori og praksis, en kløft som nogle gange synes at være ret bred.

Refleksioner over bedømmelsesprincipper og -praksis kan hjælpe lærerne med at sætte ord på deres implicitte forestillinger om og forståelse af sprogfærdighed. Det er noget lærerne kan gøre alene eller i små grupper. En del af arbejdet består i selv-analyse, men lærerne kan måske komme videre med deres refleksioner og analyser hvis de præsenterer deres tanker for en kollega eller to. Processen består af mindst to trin: beskrivelse og analyse / refleksion.

Som udgangspunkt for beskrivelsen bør læreren vælge én test- eller bedømmelsesprocedure som han eller hun har anvendt for nylig, og stille sig spørgsmål om den. Det er nyttigt at skrive svarene ned, men det er tilstrækkeligt at gøre det i skitseform; der er ikke brug for færdigpolerede formuleringer i denne fase. En grundlæggende liste over spørgsmål kunne se således ud:

- Hvad er 'sprogfærdighed' for disse sprogtilegnere i lyset af denne bedømmelsesprocedure?
- Hvordan ville du karakterisere en god præstation i forbindelse med løsningen af disse opgaver?
- Hvordan ville du karakterisere en dårlig præstation i forbindelse med løsningen af disse opgaver?
- Hvordan ser en gennemsnitspræstation ud i forbindelse med disse opgaver?
- Er det nyttigt i denne bedømmelse at overveje om de(n) bedømte sprogfærdighed(er) skal bestå af under-færdigheder? Hvis ja, hvilke?
- Hvilke andre færdigheder kræves der i bedømmelsen ud over sprogfærdigheder?

I analyse-/refleksionsfasen bearbejder man det tekstudkast der blev produceret i den foregående fase. Idéen er at fokusere på lærerens opfattelse af sprogfærdighed a) i relation til et særligt bedømmelsesinstrument, og b) generelt. Trinene er:

- Gennemgå de tekster du har produceret, og undersøg hvilke begreber du bruger til at beskrive sprogfærdighed med.
- Hvilke begreber bruger du til at skelne mellem færdigheder på et højere og et lavere niveau? Er disse begreber opgaverelaterede eller generelle (dvs. anvendelige på tværs af en række opgaver)?
- Hvilke tanker gør du dig om sprogfærdighed i denne test/bedømmelse i forhold til dine forestillinger om hvad sprogfærdighed er i mere almen forstand? (F.eks. i andre tests/bedømmelser? I undervisning generelt?)

Der er desuden behov for at analysere de filosofiske principper der ligger bag de kvalitetskriterier der bliver opstillet for forskellige slags bedømmelser. Der er allerede udført en del teoretisk arbejde inden for dette område, og det næste der er brug for, er måske analyser udført af lærere og testere af hvad dette betyder i praksis. Et godt udgangspunkt for dette arbejde er Moss' (1992) analyse af psykometriske og hermeneutiske filosofier om bedømmelse.

Psykometrisk bedømmelse

Bag traditionel testning ligger psykometriske principper. Målet i psykometrisk begrundet testning er at forudsige fremtidige præstationer så nøjagtigt som muligt. Testeren ser således efter tegn på *adfærdsmæssige regelmæssigheder* som støtte for forudsigelser. Regelmæssighed kan kun findes hvis den samme ting bliver testet mange gange og inden for en begrænset test-tid, hvilket betyder at hvert enkelt punkt der bliver testet, er ganske lille. Helheden, dvs. sprogfærdigheden, bliver så bedømt ved at lægge de små komponenter sammen. Hvis nogle komponenter er variable eller kontekstafhængige, kan de lades ude af testen fordi de ikke bidrager til afdækningen af regelmæssighed. På samme måde kan træk i præstationen som ikke optræder i hver test-tagers præstation, forbigås fordi de ikke giver et troværdigt grundlag for forudsigelser. Regelmæssighed og lighed er også værdifulde egenskaber i test-administration, ligesom sammenlignelighed bliver tillagt stor værdi. Opgaven, præstationsbetingelserne og bedømmelsesprocedurerne må være de samme for

alle test-tagere. Dette garanterer retfærdighed i målingen i psykometrisk bedømmelse.

Hermeneutisk bedømmelse

I hermeneutisk bedømmelse er det overordnede mål at forstå helheden (dvs. karakteren af en enkelt sprogtilegners sprogfærdighed) i lyset af dens enkeltdele. Fra den ene *case* til den anden bliver tolkningen og evalueringen modificeret, indtil der er taget hensyn til alle træk i præstationen. Snarere end at være en operationalisering af en teoretisk opfattelse af færdighed, er opgaverne repræsentative for den enkelte, og forskellige individer kan løse forskellige opgaver. Der bliver lagt vægt på at finde de mest passende beskrivelser af den enkeltes færdigheder, og derfor bliver rationel argumentation værdsat blandt bedømmere. Evalueringerne af dem der kender eksaminatorerne godt - det kan være lærere eller andre sprogtilegnere - bliver specielt respekteret, og begrundelsen for den hermeneutiske bedømmelses retfærdighed ligger i resultaternes alsidighed og det forsvarlige i den afsluttende individuelle bedømmelse.

Kvalitetskriterier

De kvalitetskriterier der bliver forbundet med de nævnte tilgange, er helt klart meget forskellige. Mens det måske ikke er almindeligt at finde ekstreme anvendelser af hverken den ene eller den anden filosofi i sprogundervisningens virkelighed i dag, kan det være nyttigt at holde denne skelnen mellem kvalitetskriterier for øje, især for at undgå at anvende kriterier fra den ene på eksempler fra den anden, i det mindste uden af stille spørgsmål ved hvorfor man skulle gøre det. Når vi således anvender omfattende tests der har til formål at forudsige fremtidige præstationer, vil det være klogt at sikre sig at disse instrumenter virkelig er standardiserede: troværdige, sammenlignelige, konsistente og upartiske. Dette kan ske gennem kvantitative (og nogle kvalitative) analyser.

I lyset af den aktuelle validitetsteori bør konsekvenserne af brugen af testen også tages i betragtning. I individualiseret, hermeneutisk motiveret bedømmelse er de vigtigste kvalitetskriterier gennemsigtighed og begrundelse. Eksaminanderne skal vide hvad det er der vil blive evalueret, og de skal være klar over den række af mulige opgaver de kan vælge fra. Fra evalueringsprocessen skal der være bevis på at bedømmerne anstrengte sig for at tage højde for den enkelte i tolkningen af bedømmelsen, og at den afsluttende bedømmelse kan forsvares rationelt. På grund af de sammenhænge hvor det er sandsynligt at hermeneutisk motiverede bedømmelser vil blive anvendt, og den type bedømmelsesinformation der vil blive produceret, er statistiske analyser næppe relevante for at sikre kvaliteten. I stedet kan lærerne overveje at præsentere en række bedømmelses-*cases* for en kollega, eller i nogle sammenhænge måske for eksaminanden, for at se om deres bedømmelser er forsvarlige.

Fremstillingen ovenfor giver et billede af den øjeblikkelige udvikling i de filosofiske teorier inden for bedømmelse.

Idéerne skal nu anvendes på virkelighedens bedømmelser i en uddannelsesmæssig sammenhæng for at se hvorvidt og i hvilken udstrækning de er nyttige, og hvad der yderligere er brug for for at forstå de værdier der styrer sprogbedømmelsens nyligt omdefinerede verden.

Noter

1 En tilgængelig introduktion til brugen af IRT inden for sprogtestning kan man få i Tim McNamaras bog fra 1996: *Measuring Second Language Performance.* Der er også flere introducerende lærebøger i statistik som forklarer den teori der ligger bag modellen; søgeordet *item response theory* eller *IRT* vil kunne hjælpe med at lokalisere nogle af dem i et nærliggende bibliotek. Desuden bliver de grundlæggende ting også forklaret på nogle Web-sider, som f.eks. ericae.net/scripts/cat/

2 Ét system man kan bede tilegnere om at afprøve er DIALANG, som vil blive tilgængelig gennem www.dialang.org på et stigende antal sprog i løbet af 2002. De første sprog vil være tilgængelige så snart beta-versionen teknisk set er tilstrækkelig pålidelig til offentlig testning. Flere computer-medierede evalueringsalternativer kan f.eks. findes gennem Glenn Fulchers hjemmeside Resources in Language Testing, www.surrey.ac.uk/ELI/ltr.html under 'links'-afsnittet. Hjemmesiden indeholder også et væld af andre links til andre sprogtestningsressourcer på internettet, f.eks. video-optagelser af forklaringer på grundlæggende begreber inden for sprogtestning.

3 Gode eksempler på et sådant arbejde findes i Carol Chapelles artikler (1998, 1999), som er godt læsestof for dem der arbejder med mere formelle tests som eksaminer efter afslutningen af et kursus eller som afgangseksaminer. Andre retninger inden for den standende debat fokuserer på den sociale dimension i anvendelsen af tests, og de foreslår at det er mest rimeligt at bruge sprogbedømmelse som støtte til videre sprogtilegnelse (f.eks. McNamara 2001).

Litteratur

Bachman, L. F. (2000): Modern language testing at the turn of the century: assuring that what we count counts. i: Language Testing 2000, nr. 17. S. 1-42.

Bachman, L. F. & Eignor, D. R. (1997): Recent advances in quantitative test analysis. i: Clapham, C. and Corson, D. (Eds.). S. 227-242.

Birenbaum, Menucha: Assessment 2000: Towards a pluralistic approach to assessment. i: Birenbaum and Dochy (Eds.): Alternatives in assessment of achievement, learning processes and prior knowledge. Boston; Kluwer Academic Press, 1996. s. 3-29.

Clapham, C. and Corson, D. (Eds.): Encyclopedia of language and education. Volume 7: Language testing and assessment. Dordrecht: Kluwer Academic Press, 1997.

Chapelle, C. A.: Validity in language assessment. i: Annual Review of Applied Linguistics 1999, nr. 19. S. 254-272.

Chapelle, C. A.: Construct definition and validity inquiry in SLA research. i: L. F. Bachman and A. D. Cohen (eds.): Interfaces between second language acquisition and language testing research. Cambridge: Cambridge University Press, 1998. S. 32-70.

McNamara, Tim: Language assessment as social practice: challenges for research. i: Language Testing 2001, vol. 18 (4). S. 333-349.

Moss, Pamela: Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. i: Review of Educational Research 1992, 62(3). S . 229-258.

Oversat af Michael Svendsen Pedersen