GÖTEBORGS UNIVERSITET
INSTITUTIONEN FÖR PEDAGOGIK OCH DIDAKTIK

Datorbaserade prov i engelska – en kunskapsöversikt 2009

Hans Albinsson
Institutionen för pedagogik och didaktik
Enheten för språk och litteratur

April 2009

Att bedöma språkfärdighet med hjälp av dator

Inledning

Förespråkare för datorbaserade prov framhåller ofta, förutom pappersfri distribution och automatiserad bedömning, nya möjligheter som att introducera multimedia eller färdighetssimulering, vilket skulle kunna öka autenticiteten i en provsituation. Vidare kan datorn skräddarsy ett prov för varje elev och ge omedelbar feedback, något som kan gynna lärande och lugna testtagaren.

Kritikerna menar att datormediet i sig systematiskt kan gynna respektive missgynna vissa kategorier av elever, att de testtyper som finns inte är särskilt lämpade för språkbedömning samt att kostnaderna för att konstruera och genomföra adaptiva prov är mycket stora och därmed tar resurser från andra, minst lika viktiga, områden. Säkerheten vid distribution av test där mycket står på spel, s.k. *high-stakes tests* har också debatterats flitigt.

Den första frågan man måste ställa sig i valet mellan traditionella papper-och-penna-prov och datorbaserade prov är vad syftet med en övergång till en mera maskinella bedömningsformer skulle vara. Kanske är svaret helt enkelt det att tiden är mogen och förändringen oundviklig, men det finns några ofrånkomliga ställningstaganden. Vill vi skapa ökad effektivitet eller prov som bättre reflekterar elevernas språkliga kompetenser? Hur definierar vi kommunikativ kompetens på 2000-talet? Ska proven vara i huvudsak diagnostiska verktyg eller mera tjäna som underlag för betygsättning? Och, sist men inte minst: hur kan man undvika det potentiella motsatsförhållande som stundtals tycks råda mellan effektivitet och bra bedömning, i bemärkelsen bredd och variation i relation till det som ska bedömas (*the construct*) och sätten att göra detta? ”If efficiency is the goal, the desired results are shorter, more convenient tests” (Chapelle & Douglas, 2006, s. 116). Verktygen för ”effektiva prov” finns, men de kanske inte alltid lyckas fånga de färdigheter som är relevanta att bedöma och frågan är om frestelsen att mäta det lätt mätbara någonstans är större än vid datorbaserade prov. Omfattande forskning pågår dock för att hitta bättre sätt att bedöma det som är önskvärt att bedöma, nämligen hur elever klarar autentiska språkliga utmaningar på varierande nivå. ”A test that is aiming at being a test of proficiency needs to be analytical, communicative, and integrative, among other things” (Davies, 2003)

En annan ofrånkomlig fråga handlar om vad språkfärdighet på 2000-talet innebär. Behöver grundbegreppen expanderas till att även innefatta *computer literacy*? Skulle man rent av kunna hävda att en person som inte behärskar modern teknik i vårt samhälle är så handikappad vad gäller kommunikativ kompetens att det bör påverka omdömen och betyg? Hela grundvalen för bedömning kan ifrågasättas – *a redefinition of the construct* kanske är påkallad.

Ytterligare en frågeställning är hur en eventuell övergång till datorbaserade prov påverkar lärande och undervisning, s.k. *washback*- eller *backwash*-effekter. En tänkbar följd av datorbaserade prov är att mer av lärandet förläggs till datorer, vilket på ett bättre sätt än idag skulle kunna förbereda eleverna för ett samhälle där datorn blir allt mer central som kommunikationsmedel.

Begrepp och inledande diskussion

Termerna L1, L2 och FL som används i texten står för förstaspråk, andraspråk och främmande språk.

Den i Sverige ofta använda termen ”uppgift” eller ”provuppgift” är tvetydig, eftersom den dels kan syfta på t.ex. en enskild lucka i ett lucktest, dels på hela lucktexten eller rent av hela den del av provet där lucktexten kan ingå som en del. I brist på bättre används här därför den engelska termen *item* när den enskilda luckan eller motsvarande avses.

Man brukar skilja mellan formativ och summativ bedömning (Scriven, 1968), där termen formativ står för kontinuerlig, ibland diagnostisk, bedömning, som inte är kopplad till betygsättning. Sådan bedömning brukar som regel vara *low stakes*, dvs. ingen omedelbar vinning kan uppnås genom att lyckas bra och det finns alltså heller inget incitament att försöka fuska. Med summativ bedömning brukar man bl.a. avse prov som summerar ett kunskapsområde – stort eller litet – och som stundtals är kopplad till betyg, inträde någonstans, eller till certifiering. Dessa rubriceras följaktligen ofta som *high stakes*. Termerna är inte helt självklara i alla sammanhang, och stundtals görs en dikotomisering av de båda företeelserna, som kan leda fel, t.ex. vid resonemang kring grundläggande principer för bedömning.

Det florerar ett stort antal beteckningar och förkortningar i litteraturen runt datorer och prov. Termerna *Computer-Based Testing* (CBT), *Computer-Based Assessment* (CBA), *Computer-Aided Assessment* (CAA), *Computer Assisted Language Testing* (CALT) eller *e-assessment* avser alla prov som tas vid dator, medan man med P&P menar prov som görs med hjälp av papper och penna. *Computer Adaptive Testing* (CAT) är den vanligaste termen för prov som anpassar sig efter testtagarens förmåga. Ibland läses CAA ut som *Computer Adaptive Assessment* och CALT som *Computer Adaptive Language Testing*. *Computerized Classification Testing* (CCT) innebär att testtagarna delas in i grupper beroende på resultat (t.ex. ”godkänt” eller ”icke godkänt” eller om man så vill hela betygskalan). *Web-based language assessment* (WBLA) betecknar online-prov. – I den följande texten använder vi begreppet CALT i dess mer övergripande tolkning som dator”assisterade” språkprov och termen CAT för adaptiva prov.

Computer Assisted Language Learning (CALL) står för ett system med inlärning vid datorer. Här har man ibland använt sig av något som kan rubriceras som kontinuerlig bedömning, dvs. en algoritm bedömer fortlöpande elevernas prestationer och framsteg. T.ex. har *Longman English Interactive* (Rost, 2003, refererad av Chapelle & Douglas, 2006) och On-line kurser som *Market Leader* (Longman) inbyggd bedömningsdel. Enligt Chapelle & Douglas finns i denna typ av system en stor potential att hjälpa eleverna att utvecklas i riktning mot självständighet och medvetenhet i studierna, förutsatt att kursutvecklarna har en klar uppfattning om hur bedömning påverkar lärande. (Se också avsnittet om low-stakes diagnostiska prov under för- och nackdelar med CALT/CAT.)

Bunderson, Inouye & Olsen (1988) för in ytterligare ett begrepp när de talar om de fyra generationerna av datorbaserade prov: 1. CBT 2. CAT 3. Kontinuerlig bedömning 4. IM Intelligent Measurement ”producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures”.

Linjära prov innebär att alla gör samma prov. Det finns för- och nackdelar med denna variant. De fördelar med adaptiva prov som nämns i annat sammanhang, nämligen individuell anpassning, bortfaller naturligtvis. Med tanke på att det vore svårt att genomföra datorbaserade prov, t.ex. i ett helt land, en och samma dag finns risken att innehållet i provet blir allmänt känt relativt snabbt, vilket naturligtvis skulle vara förödande vid high-stakes-prov. Fördelarna

är dock i vissa avseenden desamma som med P&P, t.ex. den pedagogiska effekt som uppnås genom att alla elever vid en gemensam provgenomgång i efterhand kan relatera till och diskutera samma uppgifter. En annan fördel är den överblick provkonstruktören får över hela provet. Vid ett adaptivt prov är det svårt att bedöma kombinatoriska effekter av de uppgifter som serveras testtagaren. Ett och samma item eller en och samma provdel uppfattas olika beroende på det sammanhang det/den sätts in i (s.k. *local dependence*). I viss mån kan man dock minska denna effekt genom att använda testlets, dvs. grupper av items, i stället för separata sådana (Chapelle & Douglas, 2006).

Sekventiella prov påminner mycket om CAT och anses väl lämpade för CCT (se ovan), medan CAT anses lämpa sig bäst för att exakt beskriva en persons färdighet. I den enklaste formen av sekventiella prov börjar testdeltagarna med samma inledande miniprov, och beroende på hur man lyckas där får man sedan ett större prov på en viss nivå (Wikström, 2005; jfr även Norges tidigare nationella prov; se nedan). En mer avancerad och flexiblare form, ibland kallad *Multistage*, innebär att *testlets* serveras; en grupp av items som hör ihop, t.ex. frågor som baseras på samma text (Way, Davis & Fitzpatrick, 2006, m.fl.). Denna variant ökar tillförlitligheten och passar t.ex. läs- och hörförståelse väldigt bra, då den är effektivare eftersom textmassan kan utnyttjas bättre, men den utgör samtidigt en utmaning för provkonstruktören; det är svårt att konstruera testlets med enhetlig svårighetsgrad. Det gäller även att tillgodose krav på variation i innehåll och i sätten att bedöma den språkförmåga som skall testas (Chapelle & Douglas, 2006). Detta ställer också krav på algoritmen eller mekanismen som väljer ut items; urvalet bör spegla en mångfald vad gäller testtyper och den språkförmåga som prövas. Om dessa överväganden överlåts åt en algoritm endast baserad på statistiska hänsyn, finns en uppenbar risk att provet blir ensidigt. Ett rimligt tillvägagångssätt skulle kunna vara att olika ”taggar” används för att beskriva dessa parametrar i en eventuell uppgiftsbank, alltså etiketter som t.ex. talar om vilket uppgiftsformatet är, om en viss, specifik färdighetsaspekt testas, vilket land en text kommer från, om inläsaren av en hörförståelsetext är manlig eller kvinnlig, har en dialektal färgning osv. Detta resonemang gäller naturligtvis oavsett om man använder sig av separata items eller testlets.

Kommer användningen av datorer i prov att påverka läromedel och undervisning (s.k. *washback* eller *backwash*)? I så fall på gott eller ont? Alderson & Wall (1992) och Wall (2000) liksom flera andra forskare diskuterar definitionen av begreppet washback, en diskussion som kanske inte ryms i denna översikt men det kan vara värt att ha i åtanke vid eventuell konstruktion av datorbaserade prov att upplägget och uppgifterna inte blir av sådan art att de inbjuder till en undervisning för prov, s.k. *teaching to the test*. Wall and Horák (2006) noterade att både lärare och elever som genomförde en TOEFL[1]-kurs upplevde att syftet med studierna var att klara provet snarare än att hjälpa eleverna att utveckla ett språk för framtiden – ett fenomen som dock inte gäller enbart datorbaserade prov.

Frågor relaterade till datormediet

Validitet

I en diskussion av datorstödd bedömning är begreppet validitet, eller giltighet, centralt. Ramarna för, och syftet med denna rapport medger dock inte någon djupare diskussion av fenomenet, dess skiftande definitioner, tolkningar och implikationer, vare sig på ett generellt vetenskapligt och/eller praktiskt plan, eller med avseende på datorstödd bedömning av

[1] TOEFL = Test of English as a Foreign Language (Educational Testing Service, USA)

språkfärdighet. Det bör dock fastslås att begreppet *construct* (i regel översatt till svenska som ’begrepp’) är centralt, nämligen att definitionen av bedömningens Vad? (och även Varför?) är av grundläggande betydelse. Som hot mot validiteten brukar anges att för lite av det definierat väsentliga tillåts ha genomslag i det som bedöms, i litteraturen refererat till som *construct under-representation,* samt att andra faktorer än dem man avser fokusera tillåts påverka resultaten, vilket brukar karakteriseras som *construct irrelevant variance* (Messick, 1989). I fallet språk kan detta t.ex. innebära att muntlig, interaktiv språkförmåga eller läsning av längre text, med syfte att fånga förmågan till inferens, ges för lite utrymme, eller att t.ex. tidsfaktorer eller tekniskt kunnande påverkar resultaten. Vidare intar *bias* i olika former en central plats i hotbilden. Senare decennier har också medfört en viss expansion av synen på validitet, som inneburit att fokus delvis flyttats från de instrument som används till en betoning även av tolkning och användning av resultat samt de konsekvenser detta får för testtagaren, men också för den analytiska process som kännetecknar utveckling och analys av bedömningsinstrument och resultat (Messick, 1989; Bachman, 2005; Kane, 2006). Vad gäller det senare finns en stor mängd litteratur, även inom området bedömning av språkfärdighet. Bland mycket annat framhålls här vikten av ett kollaborativt förhållningssätt visavi dem som på olika sätt berörs av proven och dess resultat, inte minst lärare och elever (Alderson, 1990; Bachman & Palmer, 1996; Erickson, 2009).

Alderson (1990) påpekar möjligheten att samla in värdefulla data rörande validiteten i provet genom att testtagaren får fylla i frågor (kryssfrågor eller öppna svar) om varje item; attityd till uppgiften, svårighetsgrad, vad testtagaren tror prövas, hur säker man är på svaret etc. Kanske kan man rentav genom en mer utvecklad introspektion rekonstruera de processer som leder fram till ett givet svar och på så sätt komma närmare en förståelse av hur språk lärs in. *Det kan i detta sammanhang påpekas, att systematisk, storskalig insamling och användning av testtagaren synpunkter utgör en viktig del av den utvecklings- och valideringsprocess som kännetecknar arbetet med de svenska nationella provmaterialen i engelska och övriga moderna språk (Erickson, 2009; under utgivning).*

Meunier (1994, refererad av Chapelle & Douglas 2006, s.86) talar om fem olika validitets-aspekter att ta hänsyn till vid CAT och hennes resonemang mynnar ut i en rekommendation ”to replace multiple-choice and cloze formats, and to apply the potential of CALT to live-action simulations… [to] assess students with respect to their ability to function in situations with various levels of difficulty”.

För- och nackdelar med CALT/CAT

Mycket har forskats kring skillnaden mellan CAT/CBT och P&P. Med det datorbaserade samhälle vi idag har måste man ställa sig frågan vilket som ska ses som norm, att göra prov med hjälp av datorer eller med papper och penna. Linacre (2000) menar att användningen av datorer numera minskar *test bias*, dvs. systematiskt gynnande eller missgynnande av olika kategorier provtagare. För många, kanske framför allt unga människor idag, är datorn det naturliga instrumentet för kommunikation och P&P känns väsentligt mer främmande än tangentbord och skärm. Moe, Carlsen, & Hasselgren (2006) konstaterar att många lärare har vittnat om att mindre barn har betydligt lättare att sitta still och koncentrera sig på uppgifter presenterade via dator än P&P. Att ha tillgång till hjälpmedel i datorn och online är en självklarhet för väldigt många. Om akademikers skrivförmåga skulle testas i ett prov, skulle många säkerligen uppleva att P&P innebar en orättvis nackdel (Chapelle & Douglas, 2006). Det ställs helt enkelt i dagens samhälle nya krav på kommunikativ kompetens och frågan måste ställas vilken metod som på bästa sätt återspeglar olika färdigheter och som stör minst när dessa skall bedömas. Utgångspunkten bör vara att färdigheter skall bedömas så noggrant,

exakt och verklighetstroget som möjligt och att man förutsättningslöst måste betrakta de alternativ som står till buds.

Med detta sagt kan vi konstatera att det fortfarande finns många situationer, där vi möter språk utan att datorer är inblandade. Vi pratar med andra människor, läser böcker och lyssnar till föreläsningar osv. Om man väljer att testa de färdigheter som då är aktuella via dator, är det viktigt att man försöker göra detta så autentiskt som möjligt.

Computer Adaptive Testing (CAT) innebär, i alla fall i teorin, att alla får en utmaning på rätt nivå. Duktiga elever slipper lägga tid på alldeles för enkla uppgifter och vice versa slipper svaga elever kämpa med uppgifter som är över deras förmåga. En iakttagelse som gjorts är dock att många elever upplever att alla uppgifter vid CAT-prov är svåra, och om provmodellen inte tillåter revision, d v s möjlighet att ändra tidigare givna svar, framkallar det oro. Duktiga elever är också vana vid att lugnas av ett relativt stort antal uppgifter som de upplever som lätta. I en CAT finns inga *free rides* (Chapelle & Douglas, 2006; Wandall, 2008). I Danmark har Wandall (2008) också konstaterat att CAT fungerar allra bäst på att bedöma de allra duktigaste och svagaste elevernas prestationer.

Möjligheten att gå tillbaka och revidera avgivna svar är självklar i P&P-prov och fullt möjlig i CAT, vilket visas av bl. a. Linacre (2000) och Lilley (2007). Vid första anblicken verkar revision strida mot grundtanken i CAT, och argumentet mot att ge denna möjlighet brukar vara att en testtagare, av taktiska skäl, kan svara fel på alla frågor för att få ett lätt prov och sedan vid revision svara rätt och således uppnå ett bra resultat. Linacre och Lilley visar dock att detta resonemang inte håller och att validiteten (eller ’giltigheten’) i provet inte påverkas nämnvärt av möjligheten att ändra svar. En annan faktor som skulle kunna påverka upplevelsen vid provtillfället är feedback från systemet huruvida svaren är korrekta eller inte. En undersökning visar att möjligheten att ompröva svar och feedback från systemet båda minskar provtagarens oro och förbättrar provresultatet (Chapelle & Douglas, 2006).

Datorvanan är så stor idag att farhågor om teknologins negativa påverkan på testtagarna rimligen kan anses som relativt ringa för det stora flertalet, förutsatt att testet presenteras på ett sätt som är igenkännbart för testdeltagarna. Frågan om det rent av idag är P&P-prov som ger upphov till ångest har ställts (se ovan). Naturligtvis kan man inte heller helt bortse från att det eventuellt kan finnas testdeltagare som av olika skäl inte är datorvana. Därför är det näst intill nödvändigt att övningsuppgifter finns tillgängliga, helst både före och vid själva provtillfället för att dämpa en upplevd oro och ge en viss förtrogenhet med verktygen. Wikström (2005) skriver tämligen ingående om olika aspekter av *test anxiety* och betonar vikten av övning och förprov för att dämpa nervositeten för att bedömningen ska bli så rättvisande som möjligt, i synnerhet vid komplexa uppgifter. Oro inför datorbaserade prov är ett tämligen outforskat område (Chapelle & Douglas, 2006).

Något man hela tiden måste vara medveten om är att om testtagarna av något skäl svarar annorlunda vid ett datorbaserat prov än de skulle göra vid ett P&P-prov, kommer validiteten att påverkas. (Chapelle & Douglas, 2006). Chapelle & Douglas noterar också att inte en enda studie hade genomförts fram till 2005 där betydelsen av L2-testtagares datorvana för provresultat hade prövats. 1999 gjordes dock en undersökning på TOEFL-testtagare, där en enkät kring datorvana genomfördes. Därefter utfördes ett datorbaserat och ett P&P-prov. Datorprovet föregicks av en timslång instruktion. Inga mätbara skillnader mellan resultaten från de båda distributionssätten kunde noteras.

Könsskillnader och sociala skillnader har varit föremål för omfattande studier och resultaten av denna forskning är delvis motsägelsefulla. Ripley (2008) fann inte att någotdera könet presterade sämre eller bättre när provet var datorbaserat. Inte heller Moe (2008) rapporterar några noterbara skillnader mellan pojkar och flickor vid norska nationella prov i engelska. Björnsson (2008) konstaterar att pojkar presterar bättre vid CBT och det verkar finnas en korrelation mellan textmängd och kön; flickor hanterar stora textmängder bättre än pojkar. Värt att ha i minnet kan vara att slumpval av items kan få oönskade effekter vad gäller dessa aspekter. Martin & Binkley (2009) menar också att frågan om könsrelaterade skillnader i resultat bör ses i ett vidare perspektiv, där skillnader i preferenser vad gäller informationsmetoder och -kanaler i sin helhet bör beaktas. Fördjupade analyser av datorbaserade prov i naturvetenskap visar dock på vissa systematiska differenser mellan pojkars och flickors resultat på datorbaserade respektive P&P-bjudna prov, vilket av vissa forskare påtalas som ett potentiellt validitetshot, som kräver vidare analyser och ställningstaganden (Halldórsson, McKelvie & Björnsson, 2009; Sørensen & Møller Andersen, 2009).

Vid utprövning och kalibrering av items är det viktigt att vara medveten om den språkliga och kulturella miljö inom vilken utprövningen görs. Kulturella skillnader och språklig bakgrund påverkar lösningsfrekvensen (Brown and Iwashita, 1996). Även ovana vid vissa typer av provuppgifter kan påverka resultatet. Detta har bland annat konstaterats i EU-projektet EBAFLS (Building a European Item Bank of Foreign Language Skills)[2].

Vad gäller reliabilitet vid datorbaserade prov kan man peka på att ett adaptivt prov i alla fall i teorin fortsätter att mäta till dess en reliabel bedömning har gjorts. På plussidan kan också nämnas möjligheten att ge *partial credit* (se uppgiftstyper nedan) och att datorn är konsekvent i sin bedömning. En gammal sanning vid P&P-prov, nämligen att ju längre provet är desto högre blir reliabiliteten, stämmer inte vid CAT, utan i stället konstaterar Chapelle & Douglas (2006) att ”the better the items fit the examinee, the higher the reliability”.

Datorskärmen kan introducera oönskade effekter, främst vid läsning av text. Även om dagens skärmar generellt är av god kvalitet kan man inte helt bortse från att olika skärmar skapar olika premisser för testtagarna. CRT (gamla ”bulliga” monitorer) och TFT/LCD (platta skärmar) lär finnas sida vid sida några år till och de kan återge text något olika. Överhuvudtaget är layouten mycket viktig; var bilder och knappar osv. placeras kan ha avgörande betydelse för upplevelsen av provet. Val av typsnitt vid provets utformning kan vara av stor vikt för hur lättläst texten upplevs. Ska fonten t.ex. vara punkt- eller vektorbaserad? Kommer användarens webbläsare att ersätta inlagda typsnitt? Skärmupplösningen och därmed typsnittens läsbarhet och den textmassa som får plats på skärmen varierar troligen mycket mellan olika arbetsstationer. Val av webbläsare och operativsystem kan ge olika förutsättningar för olika testtagare. Även om gängse bruk i sammanhanget är att ange minimikrav på upplösning kan problem uppstå vid längre sammanhängande text, något som kan vara aktuellt på högre stadier. Användaren tvingas då skrolla eller flytta sig fram och tillbaka mellan sidor, vilket sannolikt försvårar läsningen jämfört med ett P&P-prov. Haas & Hayes rapporterade (1986, citerad av Clariana and Wallace, 2002) att om en textpassage som förknippas med en uppgift sträcker sig över mer än en sida blir resultatet sämre vid ett datorbaserat prov än vid P&P. Bridgeman (2008) konstaterar att testtagare som har tillgång till stora skärmar med hög upplösning når bättre resultat på läsförståelse än de som har små skärmar med låg upplösning; skrollning är negativt för testresultatet. Vid miniutprövningar (s.k. *piloteringssesjoner*) i Norge konstaterade Moe, Carlsen, & Hasselgren (2006) samma sak och som ett resultat

[2] http://www.cito.com/research_and_development/participation_international_research/ebafls.aspx

förekommer inga läsförståelseuppgifter där skrollning är nödvändig i årskurs 5. Om testdeltagarna är äldre och bär progressiva glasögon, kan det i ogynnsamma fall ytterligare försvåra läsningen på skärm och ge en onaturlig och tröttande ställning för huvudet. Det kan således bli svårt att se uppgifter och text samtidigt för vissa testtagare. Om inte sådan funktionalitet byggs in i verktygen som användaren har att tillgå, kan man inte heller stryka under, dra linjer mellan olika textelement eller göra anteckningar, strategier som får anses normala vid textläsning.

Reaktioner från brukarna kan på olika sätt påverka utfallet vid datorbaserade prov. Lärare är fortfarande i stor utsträckning osäkra inför den nya tekniken och kan reagera negativt på formatet. Det behöver undersökas om detta är något som kan färga av sig på eleverna och kanske kan innebära att visa elever blir sämre förberedda för datorbaserade prov. CAT ger som regel korta prov, något som kan kännas utmanande och testtagarna kan uppleva frustration att de inte får visa vad de kan (Tonidandel & Quiñones, 2002). Omvänt kan elever som får många fler provuppgifter än eleverna vid angränsande datorer bli osäkra och kanske dessutom störda. (Denna aspekt behöver antagligen ägnas uppmärksamhet vid eventuella datorbaserade prov. Ska elever som är färdiga lämna rummet eller ska de sitta kvar och vad gör de i så fall? Se även Danmark nedan under rubriken CALT i världen.) Ytterligare en faktor som inte ska underskattas i detta sammanhang är att testtagare kan reagera på gränssnittet och göra dåligt ifrån sig bara av det skälet (Magal-Royo, 2007). Det finns en del exempel på tilltalande uppgifter men betydligt fler på oinspirerande, för att inte säga fula, layouter.

Den tid som avsätts för att göra ett CAT-prov får inte vara för snålt tilltagen (Bridgeman, 2008). Det blir ett rättviseproblem, eftersom alla items inte tar lika lång tid att svara på och vid adaptiva prov testtagarna dessutom kan/bör få olika många uppgifter. Dessutom klarar algoritmerna inte av att hantera de slumpvisa gissningar som elever med tidspress ofrånkomligen kommer att producera.

Forskningen kring datorbaserade prov har i stor utsträckning fokuserat på effektivitetsvinster och väldigt få studier har bedrivits rörande hur datorn skulle kunna bedöma på sätt som traditionella test inte kan (Chapelle & Douglas, 2006).

Tiden det tar för testtagaren att besvara uppgifter kan mätas och datorn kan logga och tolka testtagarens väg genom provet. Speciellt det senare skulle kunna vara intressant om online-hjälp finns tillgänglig. I vissa testmodeller finns t.ex. enspråkiga lexikon med utvalda glosor från den aktuella texten tillgängliga (se t.ex. Dialang[3]) och då skulle man kunna utvärdera testtagarens sätt att utnyttja hjälpen. Hur tillgång till online-hjälp påverkar validiteten är dock relativt outforskat (Chapelle & Douglas, 2006). Huruvida detta är relevant lämnar vi därhän, men det framhålls ibland som en fördel med datorbaserade prov. Alderson (1990) pekar på att datorn kan ge omedelbar feedback till testtagaren om avgivna svar är korrekta eller inte och ge en ny chans att svara. Ett korrekt svar vid andra försöket skulle kunna ge reducerad poäng, något som påminner om det system med *partial credit* som diskuteras på sidorna 7 och 16. Alderson påpekar också beträffande online-hjälpmedel att testtagaren själv kan ges möjlighet att fatta beslut om huruvida hjälpen behövs eller ej, och att hjälpmedlen kan göras tillgängliga för utvalda testtagare med särskilda behov.

De CAT-baserade nationella provsystem som finns runt om i världen använder sig av en

[3] Dialang = Diagnosis of language ability (diagnostiska prov i 14 språk, framtagna med stöd av EU-kommissionen)

fastlagd tidsperiod snarare än en gemensam dag för provens genomförande. Om proven distribueras från en central server är detta närmast nödvändigt med hänsyn tagen till belastningen, i synnerhet om man väljer att använda multimedia i provet.

Att konstruera, upprätthålla och distribuera en CAT är mycket dyrt. Alla ingående delar och steg i processen kräver kompetent expertis och en stor provbank av delvis ny typ måste skapas. En CAT kräver att det finns uppgifter av drastiskt olika svårighetsgrad, från mycket lätta till mycket svåra och banken måste vara tillräckligt stor för att tillförlitligheten skall kunna garanteras. Alla dessa bitar kräver mycket stora finansiella resurser, främst i uppbyggnadsskedet, men även kontinuerligt för att upprätthålla kvalitén. Chapelle & Douglas (2006) noterar att det typiska scenariot när man ger sig in på ett teknologibaserat projekt är att det tar längre tid och kostar mer än man trodde och att man ofta börjar utan att ha tillgång till nödvändig expertis, eftersom sådan är svår att uppbringa. I Dialang-projektet (se CALT i världen nedan) lades *item-level adaptive tests* ner pga. svårigheten att konstruera item-banken. Detta var också en bidragande orsak till att TOEFL frångick adaptiva prov (se TOEFL under rubriken ”CALT i världen” nedan). “We have learned that computer adaptive testing is hideously expensive in large-scale tests. Item pools became vats, then lakes, then oceans, just to maintain test security in environments like China. The era of adaptivity in mass international language testing is dead.”[4]

Datorbaserade low-stakes diagnostiska prov skulle eventuellt kunna vara ett verktyg för lärare att tidigare upptäcka problem och göra det möjligt att sätta in rätt stödåtgärder. Alderson (1990) diskuterar skillnaderna mellan prov och övning och konstaterar att datorer möjliggör grundlig analys av elevers svaga och starka kunskapssidor. Han skissar ett scenario där eleven under pågående prov slussas över till en övningsdel för att arbeta med eventuella problem och sedan återgår till provdelen när programvaran bedömer att momentet behärskas, eller också låter man testa elevens brister på det aktuella området grundligare genom att fler och mer detaljerade uppgifter ges. Svårigheten, menar Alderson, ligger här inte i otillräckligheter i hård- eller mjukvara, utan i våra bristande insikter om hur språk lärs in, varför olika individer gör olika fel, hur god språkundervisning ska se ut osv. Det finns också ett amerikanskt system, NWEA MAP, som använder CAT för att diagnostisera elevers nivå i olika ämnen och sedan föreslår strategier för att arbeta vidare för att uppnå en förbättring. Bl.a. staten Iowa använder sig av NWEA-lösningar för sitt ”nationella” testsystem (NWEA).

En odiskutabel fördel med datordistribution av prov är möjligheten att göra kontinuerlig utprövning av nya items eller provdelar som en del av ordinarie prov. Dessa items räknas då inte in i resultatet, utan syftet är enbart att bedöma validiteten i uppgifterna. Vinsterna med detta är flera; testtagaren vet inte att en uppgift ”bara” är för utprövning och anstränger sig därför förmodligen mer än annars, utfallet kan relateras till elevens prestation på provet i sin helhet (i stället för som nu är brukligt, till en provdel med jämförelsevis få poäng), det relativt tidskrävande arbetet att hitta skolor och lärare som är villiga att genomföra utprövningar bortfaller och alla data kommer direkt i elektronisk form, klara att bearbetas.

Som framgår av ovanstående är vissa aspekter av datorbaserade prov tämligen noggrant undersökta, medan andra har ägnats mycket lite uppmärksamhet. Behovet av fortsatt forskning är alltså uppenbart.

En bra checklista vid ev. övergång till datorbaserade prov är sammanställd av Assessing

[4] Better communications test will silence critics. Glenn Fulcher, guardian.co.uk, Friday 18 November 2005, http://www.guardian.co.uk/education/2005/nov/18/tefl3

Learning in Australian Universities (se denna referens).

Frågor relaterade till bedömningsfokus

Produktiva färdigheter

Skriva

En mycket vanlig variant är att ett prov skrivs på dator och bedöms manuellt, men varianter för maskinbedömning finns. Den mest använda är e-rater som används av flera testorganisationer, bl. a. TOEFL (sedan 2005 använder man dock inte dator för bedömning av denna färdighet) och tjänsten Criterion. E-rater, som är baserad på en teknik kallad NLP (*Natural Language Processing*), kan analysera strukturen i det skrivna (inledning, avhandling avslutning), hur avancerat ordförrådet är, ordförråd som är relevant för ämnet, syntax, stavning, satsstruktur, variation i meningsbyggnad och interpunktion. Hög korrelation med mänskliga bedömare har kunnat konstateras på de flesta punkter (Lee, Gentile & Kantor, 2008), men den kanske viktigaste aspekten, nämligen hur en text uppfattas av en läsare, kan datorn inte bedöma fullt ut (Bridgeman, 2008). Texter som har skrivits för att få systemet att misslyckas har demonstrerat detta (Chapelle & Douglas, 2006). Manuell medbedömning sker kontinuerligt på ett slumpurval av de datorbedömda arbetena och på grundval av de diskrepanser man upptäcker förbättras algoritmen fortlöpande. E-rater är endast utvecklad för L1 och är sannolikt inte lämplig för L2/FL, då bedömningen av L2/FL-produktion, i alla fall delvis, styrs av andra kriterier.

Om Criterion: Feedback ges i form av påpekanden om fel och tips om vad som kan förbättras. Viktigt är att påpeka att programmet ger en lista på ämnen att skriva om; det är alltså inte generellt användbart för all skriftlig produktion. (Möjligheten finns att använda detta som en automatiserad tjänst där elever eller studenter får en chans att arbeta om innan ett arbete lämnas in. Man kan bestämma hur många gånger omarbetning får ske innan inlämningen görs.)

Ett program som liknar e-rater till sin funktion är Intellimetric från Vantage Learning. Det vilar på vad man kallar “a "brain-based" or "mind-based" model of information processing and understanding”. Beträffande problemet som nämndes ovan med e-rater och hur programmet hanterar nonsens med rätt ord och grammatik, sägs Intellimetric klara 95 % av sådana försök. Det används bl. a. i. GMAT (se d.o. nedan; Intellimetric).

För att kunna bedöma färdigheten *skriva* för L2 har ett intressant försök gjorts med att lista felen i ett utprövningsmaterial, 480 TOEFL-uppsatser, skrivna av elever med olika L1 (modersmål) och sedan använda resultatet dels för att katalogisera vilka typiska fel olika grupper av L1-användare gör, dels för att försöka skapa en algoritm som rankar dessa misstag. Detta kanske kan vara ett sätt att komma vidare med automatisk bedömning av färdigheten skriva hos L2- och FL-elever. (Lee, 2007). Hinkel har listat ordval hos FL (elever som lär sig engelska) som bidrar till uppfattningen att språket är torftigt och inexakt då de skriver, något som också skulle kunna ingå i en framtida bedömningsmodell för maskinell bedömning av testtagare som lär sig ett främmande språk (Hinkel, 2003, refererad av Chapelle & Douglas, 2006).

Med all sannolikhet kommer de närmaste åren att se stora framsteg inom detta område. Genom en förväntad utveckling inom datortekniken i allmänhet och artificiell intelligens i

synnerhet, kan man i framtiden räkna med en mer ”intelligent bedömning” (Bennet, 1999 refererad av Chapelle & Douglas, 2006)

Olika personer skriver olika snabbt. Frågan om sämre tangentbordskompetens skall tillåtas påverka provresultatet hänger samman med frågan hur man definierar kommunikativ kompetens idag. Bridgeman (2008) konstaterar att bristande tangentbordskompetens kan vara negativt för testtagaren och att olika tangentbord kan påverka; att skriva på en bärbar dators kompakta tangentbord ger t.ex. en nackdel, jämfört med ett fullstort.

Rättstavningshjälp är en annan parameter som bör utredas. Är det kanske så att nästan allt som skrivs nu produceras i en ordbehandlare och att det till kompetenserna idag hör att kunna utnyttja rättstavning och datorbaserade lexikon?

Tala

Ofta använder vi rubriken ”tala och samtala”, och när vi gör det vidgar vi begreppet till att gälla interaktion mellan två eller flera parter. Detta samspel kan datorn idag, så vitt vi vet, inte bedöma alls. Inte heller kan en maskin tolka och ”förstå” mänskligt tal. De problem som diskuteras ovan under färdigheten skriva och nedan kring bedömningen av öppna svar i läs- och hörförståelseprov gäller naturligtvis även vid talat språk och till detta kommer svårigheten att överföra talet till ett för datorn begripligt format. Den största delen av forskningen och utvecklingen på detta område ägnas åt just detta steg. Vad kan då datorn göra?

Molholt och Presler konstaterade i en pilotstudie 1986 att datorer är bra på att bedöma uttal. Hög korrelation med mänskliga bedömare uppmättes vad beträffar saknade fonem (språkljud), fel fonem, partiella fonem, extra fonem och betoningsfel. Andra studier uppvisade inte samma positiva resultat, t.ex. Reid, samma år (refererad av Chapelle & Douglas, 2006). Gemensamt för de maskinbedömda modellerna är att yttre form bedöms, men hög korrelation med mänskliga bedömare har i de flesta fall uppmätts (Chapelle & Douglas, 2006).

Versant test (tidigare PhonePass/SET-10) Ett prov konstruerat för FL. Korrekthet vid repetition av ord bedöms, liksom uttal, flyt vid uppläsning av text, upprepning av ord och frågor av typen ”välj alternativ A eller B”. Ytterligare en del som inte bedöms utan sparas så att behöriga personer kan lyssna innehåller frågor av typen ”What qualities do you look for in a friend?” Testtagaren har sedan 20 sekunder på sig att formulera ett svar. Programmet använder sig av röstigenkänningsprogramvara som innehåller en algoritm baserad på ett stort bibliotek av röster tillhörande infödda talare med olika regional och social tillhörighet. (Chapelle & Douglas, 2006; Versant). Chun (2006) riktar kraftig kritik mot detta test, som han menar saknar autenticitet (”it fails by any reasonable measure of authenticity”) och därmed validitet.

SpeechRater från ETS är ett relativt nytt program som sägs kunna tolka fritt tal från FL/L2-talare. TOEFL använde sig tidigare av denna teknik men har nu gått tillbaka till mänskliga bedömare. Det hävdas att uttal, *fluency*, grammatisk korrekthet och vokabulär kan bedömas. Systemet klarar i dagsläget inte att bedöma innehållet i det som sägs och går att lura genom att man läser innantill i något som inte har det minsta med det aktuella ämnet att göra. Det är med andra ord inte redo för *high-stakes* testning (SpeechRater).

En variant som förutsätter att programmet lär sig varje talare är DragonNaturallySpeaking från Dragon Systems. En text bestående av ca 3800 ord läses in för att programmet ska lära

sig talaren och sedan läses den egentliga provtexten upp, ca 1050 ord. Försöket har gett resultat som är jämna och konsekventa.

POET/OEPT Purdue University har (under ledning av professor April Ginther) utvecklat ett eget prov för färdigheten ”tala”. Systemet arbetar med mänskliga bedömare och benchmarks som presenteras på skärmen. (POET/OEPT, Chapelle & Douglas, 2006)

COPI Computerized Oral Proficiency Instrument från Center for Applied Linguistics. Mänskliga bedömare lyssnar online på testtagarna och resultatet sparas och finns tillgängligt.

Röstigenkänning bygger på att programmet har tillgång till en stor ordlista. Ord som inte finns i ordlistan riskerar att tolkas som feluttal av ord som finns listade. (Burstein, Kaplan, Rohen-Wolff, Zuckerman & Lu, 2000)

I sammanhanget kan också nämnas den variant Dialang skissar på i sina nya experiment-uppgifter. De kallar den ”indirekt tal med ljudalternativ”. Det är *MCT*, där ”talaren”, d v s testtagaren väljer svar bland fyra ljudklipp och provtypen kanske snarast ska ses som hörförståelse.

Framtidsperspektivet är intressant. Utvecklingen går framåt, både vad gäller röstigenkänning och talsyntes (se ovan). Kanske blir det en dag möjligt att koppla samman röstigenkänning, analys genom artificiell intelligens och talsyntes och på så vis skapa interaktiva talsituationer med hög autenticitet, och kanske vågar man till och med hoppas att datorn då förmår bedöma detta på ett relevant sätt. Men vägen dit är fortfarande oerhört lång.

Receptiva färdigheter

Diskussionen om receptiva färdigheter har som utgångspunkt de uppgiftstyper som normalt återfinns i datorbaserade prov. Dock är uppgiftstyper kanske en förlegad utgångspunkt. Man utgår från datorns begränsningar i stället för att utgå från konkreta språksituationer som är meningsfulla att testa och sedan undersöka om datorn är ett lämpligt instrument för denna prövning. Trots detta känns en inventering av det som finns idag meningsfull. Enligt Chapelle & Douglas är valet av uppgiftstyper, utformningen av dem och det sätt på vilket vi utvärderar svaren i relation till det vi önskar bedöma, avgörande för kvaliteten i bedömningen (2006). En grundläggande distinktion vid diskussion kring svarsformat är om det handlar om *selected response* (slutna svarsformat), där testtagaren väljer mellan ett antal möjliga, givna svar eller *constructed response* (öppna svar), där han eller hon själv skriver svaret.

Flervalsfrågor (*MCT*) är sannolikt den vanligast förekommande uppgiftstypen vid dator-baserade prov. Många anser att den är mindre lämplig än former av öppna svar, därför att den dels har relativt hög ”gissningsfaktor”, dels öppnar för teststrategier som att utesluta orimliga alternativ och att genomskåda testkonstruktörens intentioner, vilket i sin tur negativt påverkar *construct validity*. Flera av nedanstående uppgiftstyper (*drag-and-drop*, *hot spot* etc.) är strängt taget MCT, eftersom det finns en begränsning i antalet möjliga svar, men då antalet möjliga svar oftast är mycket stort torde dessa typer inte uppvisa samma brister som tradi-tionell MCT. En term man ibland stöter på för dessa typer är *complex MCT*. (Exempel: Dialang)

Sant/Falskt är den enklaste typen av flerval. Denna typ kräver på grund av slumpfaktorn ett mycket stort antal items för att en säker bedömning skall kunna hävdas.

Matching är en uppgiftstyp där man kombinerar t.ex. uppgifter från lista A och lista B, rubriker med texter, eller definitioner/beskrivningar med textinnehåll. Vid datorbaserade prov används ofta drag-and-drop vid matching. (Exempel: Dialang)

Drag-and-drop (*DnD*) innebär att objekt dras med musen till rätt plats, t.ex. en bild.

Hot Spot är en eftersökt punkt eller ett område. Testtagaren klickar exempelvis på något ställe på en bild eller markerar ett ord, en mening, en rad eller ett stycke i en textmassa. (Exempel: Dialang)

Tematisk gruppering är en form av MCT där testtagaren ombeds klick-markera alla ord i en lista som har något gemensamt, t.ex. ord för träd, blommor eller släktskap osv. (Exempel: Dialang)

Deletion innebär att testtagaren markerar det överflödiga ordet i en sats. En typ som lämpar sig i sammanhang där man t.ex. vill testa grammatisk korrekthet. (Exempel: Dialang)

Insertion är motsatsen till deletion, dvs. ett ord fattas och med hot spot-teknik markerar testtagaren var i satsen ordet ska sättas in. (Exempel: Dialang)

Öppna svar är en vanligt förekommande uppgiftstyp i svenska prov. De främsta skälen till detta är att den anses fokusera och kunna bedöma kunskap på hög kognitiv nivå och man undviker de ovan nämnda nackdelarna med *MCT*. Bedömningen av öppna svar kan dock skapa problem vid maskinell granskning. *One word gaps* innebär att det sökta svaret består av endast ett ord och är den variant av öppna svar som en dator har lättast att ta ställning till, förutsatt att antalet acceptabla svar är begränsat. Om felstavade men fullt begripliga svar ska accepteras, måste antingen alla felstavningsalternativ som bedöms kunna förekomma som svar och som anses vara acceptabla läggas in i mallen, eller också får man söka en lösning med jokertecken (*wild cards*), där t.ex. tre godtyckliga bokstäver tillåts avvika från den tänkta lösningen. Båda kan naturligtvis vara vanskliga metoder, speciellt vid high stakes CALT-prov. Det är mycket svårt att förutse alla stavningsvarianter som kan förekomma och jokertecken kan ge poäng för felsvar, i synnerhet vid ord där en enda utbytt bokstav kan ge en helt annan betydelse. Om man använder sig av den förstnämnda varianten vid t.ex. ett luck-test, kan många korrekta alternativ finnas för varje lucka vilket kan ge en mycket omfattande mall om man ska lista alla felstavningsalternativ.

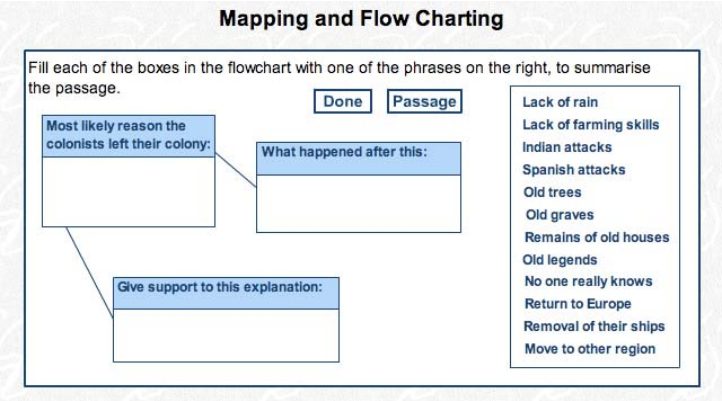
Beträffande öppna svar bestående av fler ord eller hela meningar kan sägas att det har varit svårt att hitta dokumentation om hur detta i praktiken är möjligt att bedöma maskinellt. Testmodellen förekommer t.ex. i det WebLAS-exempel med en videoinspelad akademisk föreläsning som nämns nedan (WebLAS). (Den troliga förklaringen till att det är svårt att hitta dokumentation är att det ligger stora ekonomiska intressen i de system som tagits fram och att man därför vill slå vakt om de framsteg man gjort. Det verkar delvis vara en trend att mer och mer forskning kring datorbaserade prov flyttas från universitet till privata företag, och detta påverkar naturligtvis öppenheten.) En tänkbar modell är att frågan snävar in det möjliga svaret så att få korrekta varianter kan tänkas och att man sedan gör en mall på samma vis som beskrevs ovan, med alla poänggivande svar listade. En annan modell, som skulle kunna fungera, är att man listar nyckelord som skall finnas i svaret, men en förutsägbar svårighet med båda dessa förfaringssätt är språkets inneboende rika variation och flexibilitet; det är helt enkelt väldigt svårt även efter en mycket omfattande utprövning att förutse alla tänkbara varianter. Mer komplicerade algoritmiska lösningar används troligen också.

”Anledningen till att det är så svårt att datorisera öppna uppgifter är att rättningsalgoritmen som skall konstrueras är oerhört komplicerad, och öppningar för felkällor i tolkning och rättning mycket stora. Detta gör att rättningen i regel blir kostsammare, mer tidskrävande och med större risker för fel än vid traditionell rättning” (Wikström, 2005). Chapelle & Douglas (2006, s. 94) uttalar sig försiktigt optimistiskt: ”The use of computer algorithms to score such responses may make this response format usable, thereby increasing the potential for construct validity”. Den princip som i regel tillämpas i svenska nationella prov i främmande språk, nämligen att alla svar som är (korrekta och) begripliga för en infödd talare skall ge poäng försvårar sannolikt ytterligare användningen av öppna svar i CALT. Ett vanligt förfaringssätt vid datorbaserade prov är att öppna svar sållas ut och bedöms manuellt, men denna metod är naturligtvis inte förenlig med CAT.

En ofta förekommande bedömningsmodell vid öppna svar är *partial credit*, dvs. att helt korrekt svar ger full poäng och att avdrag görs för avsaknad av information eller fel av olika slag. Då detta anses bättre återspegla testtagarens kompetens, bedöms det ge bättre reliabilitet (Chapelle & Douglas, 2006).

En möjlig läsförståelseuppgift består i att en text styckas upp i ett antal delar som presenteras i slumpvis ordning och testtagaren ombeds sedan arrangera elementen så att ursprungstexten återskapas. Den engelska termen är *re-organisation*. Testtypen är vanlig t.ex. i Cambridgeprov och kan ge upphov till problem med poängsättning. Att ge en poäng för helt rätt lösning och noll för alla fel speglar kanske inte på ett korrekt sätt språkförmågan, då en elev som har rätt inbördes ordning på flera element men inte alla kan antas ha bättre textförståelse än den elev som inte lyckas alls. Även här används lämpligen ett system med *partial credit*, dvs. att olika poäng eller delar av poäng ges för svar som bedöms som delvis rätt (Chapelle & Douglas, 2006).

En typ av *Drag-and-drop matching*, kallad *Mapping and Flow Charting* från Dialangs nya experimentexempel ses här. Fraserna till höger dras till rätt box efter det att testtagaren läst texten som frågorna baseras på. Knappen *Passage* leder tillbaka till texten. Respons på svaret (t.ex. *2/3 correct*) ges när man tryckt på knappen *Done* och om svaret inte är korrekt har man möjlighet att försöka igen. (Exempel: Dialang)



Färdighetssimulering, typ flygsimulator. Här talar man om *interactive authenticity*. Kyllonen (2008) presenterar enkla modeller för simulering. En tänkbar uppgift t.ex. i tyska: På hemsidan db.de (Deutsche Bundesbahn) tar testtagaren fram förbindelser och tider till viss destination med vissa kriterier uppfyllda. Denna provtyp anses bedöma färdigheter på en hög kognitiv nivå på ett sätt som upplevs autentiskt. Det anses vara motivationshöjande för tekniskt intresserade, avskräckande för andra. Provångest kan förstärkas hos dem som lider av det, oftast kvinnor, enligt Moe & Johnson refererade av Linacre, 2000. Färdighetssimulering är en omdebatterad och relativt outforskad provform. Kritiker menar att den är så högt korrelerad med MCT att traditionella typer av prov estimerar förmågan bättre (Wikström, 2005).

(En intressant tanke är att utnyttja kompetens inom interaktiva dataspel för att skapa autentiska språksituationer, kanske en virtuell värld av typen *Second Life*. Som sagt, spännande, men detta är outforskad mark och antagligen är det inte genomförbart idag.)

Datorn skulle kunna ge testtagaren möjlighet att använda inbyggda hjälpfunktioner eller backa och lyssna om vid hörförståelse om det bedöms som önskvärt. Man kan, som tidigare nämnts, tänka sig att logga och utvärdera strategier för att klara uppgiften, vilket skulle ge en helt ny typ av bedömning (Chapelle & Douglas, 2006), en *redefinition of the construct*. Detta område behöver dock utforskas närmare, eftersom det på ett genomgripande sätt förändrar hela bedömningen.

Multimedia i hörförståelse förtjänar särskild uppmärksamhet eftersom detta oftast nämns som en fördel med CBT/CAT, främst därför att det anses skapa *situational authenticity*. Dialang har ett enkelt exempel på video i hörförståelse (Dialang), WebLAS en mer avancerad; en akademisk föreläsning i psykologi om minnet spelas upp och därpå följer frågor (WebLAS). Coniam (2001 refererad av Chapelle & Douglas) menar att om hörförståelse ska bedömas kan videoinslag riskera validiteten hos provet. Han fann å andra sidan inga noterbara skillnader då han jämförde ren lyssning med videostödd lyssning. Ockey (2007) noterar också att introduktion av video i hörförståelse ökar autenticiteten, men samtidigt förändrar *the construct*. Provdeltagarna tolkade gester och ansiktsuttryck och läste på läpparna. Förmågan att dra nytta av detta varierade från en testtagare till en annan. Personerna i undersökningen fick också ta ställning till om videostödet var till nytta för förståelsen eller distraherande. Svaren varierade mellan testtagarna. Ockeys testunderlag var mycket litet (sex personer), men hans slutsatser känns ändå relevanta. Wilhelm & Schroeders (2008) fann att lyssningsprov med bildstöd fungerade mycket bra i ett försök med FL (engelska). De fann också att lyssningsprov på dator gav bättre bedömning än traditionella Hf. ”Aural presentation of stimuli and responses is obviously more adequate.” Påpekas kan dock att de exempel på svarstyper de visar upp alla är MCT.

Vad gäller autenticitetsdiskussionen är det värt att poängtera att ren hörförståelse (*receptive listening;* Buck, 2001) sällan förekommer i verkligheten; man ser oftast den som talar eller också har man bildstöd som illustrerar det som sägs, t.ex. vid en nyhetssändning. Unga människor lyssnar inte i större omfattning på tal-radio, vid telefonsamtal kan man be om repetition osv.

Lyssningsprov vid dator förutsätter sannolikt hörlurar. Detta kan möjligen, i alla fall initialt, vara ett problem på skolorna, men kan samtidigt ses som något som skulle kunna förbättra likvärdigheten i bedömningen, eftersom man i viss mån kommer ifrån problem med skiftande akustik i olika lyssningslokaler och på olika platser i samma lokal. På senare tid har hörlurar som stänger ute omgivningsljud på elektronisk väg blivit allt vanligare och därmed överkomliga i pris. Dessa skulle sannolikt vara ytterst lämpliga i detta sammanhang. Bl.a. TOEFL använder sig av denna typ.

Självbedömning

Datorer torde lämpa sig mycket väl för självbedömning. Dialang har t.ex. experimentmodeller där man lyssnar till exempelpersoner eller skriver en text, jämför med färdiga, bedömda mallexempel och tar ställning till om man själv kan lösa uppgiften bättre eller inte, s.k. *benchmarking*. Renets prov (se ovan) börjar med en självbedömningsdel. Kamratbedömning borde också vara fullt genomförbar och skulle då t.ex. kunna ingå i en datorbaserad portfolio.

En intressant variant av självbedömning är en modell som kombineras med uppgifterna. Testtagaren svarar på provuppgiften och kryssar sedan i en ruta där man bedömer i hur hög grad han/hon är säker på det avgivna svaret (*confidence in chosen response*; exempel: Dialang). En sådan modell skulle kunna ge mycket värdefull information till både eleven och läraren om hur eleven uppfattar sitt kunnande i relation till den faktiska förmågan, något som skulle kunna leda till fruktbara ämnessamtal.

CALT i världen

Datorbaserade språkprov används på många håll i världen, både i statligt finansierade program, universitetsprojekt och i mer kommersiella tillämpningar. Här följer en lista på några av dem.

Nationella prov

Norge har datorbaserade nationella prov i bl. a. engelska. Pisa-undersökningen 2000 kom som en obehaglig överraskning för många i Norge. Det visade sig att resultaten, mot allmän förväntan, enbart låg strax över medel; landet var klart distanserat av Sverige och långt bakom Finland. För att komma tillrätta med problemet lanserades ett nationellt provsystem, där färdigheterna läsa och skriva (norska), matematik och engelska (läsa och skriva) ingick. Proven är *low stakes* för testtagaren och syftet är att ge elever, lärare och skolledare nödvändig information för att befrämja pedagogisk utveckling, samt att förse lokala och nationella myndigheter och allmänheten med information som kan stimulera till dialog och utveckling av utbildningsstandarden. Det pedagogiska syftet togs dock bort i den förändring 2006 som beskrivs nedan.

Provutvecklingen började år 2003. Under 2004 genomförde man storskaliga utprövningar över hela landet och under 2004 och 2005 gjordes proven av elever i årskurserna 4, 7, 10 och 11. Det engelska provet var datorbaserat; de övriga P&P. År 2006 låg provverksamheten nere och vissa förändringar gjordes; man valde att avstå från färdigheten skriva eftersom interbedömarreliabiliteten ansågs för låg, gymnasieproven lyftes bort, delvis på grund av kraftiga protester mot centraliserade prov (alltså inte mot datorbaserade prov) från gymnasieelevernas sida. Övriga prov flyttades från vår (åk 4, 7) till höst (åk 5, 8) men bedömer kraven för 4 och 7 (en sommar emellan). Anledningen till denna förändring var att lärarna kände att även deras arbete utvärderades och eftersom eleverna byter stadium och får nya lärare efter dessa årskurser blev detta mindre känsligt. Även de engelska provens karaktär förändrades något. I sin senaste utformning använder man sig av tre likvärdiga linjära prov (med exempeluppgifter åtkomliga via Internet). I den första versionen hade man en typ av sekventiella prov, där ett inledande prov gav testtagaren ett av tre möjliga linjära prov med olika svårighetsgrad. 60 000 elever deltar i vardera åk 5 och åk 8. Allt relateras till CEFR A1-B1 i åk 5, A1-B2 i åk 8. Vokabulär och grammatik planeras ingå i proven från och med 2009 och senare även färdigheten *höra*. Utprövningen görs i datormiljö, där alla items utprövas i olika kombinationer och av detta sätts tre likvärdiga prov samman. Problemet som uppstod när man använde adaptiva/sekventiella prov var att den norska skolmyndigheten ville ha poäng och poängen i proven var inte jämförbara, vilket skapade stora svårigheter. Att provresultaten offentliggjordes väckte också starka reaktioner, men proven som sådana hade och har fortfarande stor acceptans och numera anses publiceringen positiv; den leder på många håll till frågan ”varför?”, vilket i förlängningen kan initiera en behövlig förändringsprocess. Lärarna i engelska var mycket nöjda med att de, tack vare den datorbaserade rättningen, inte fick en ökad arbetsbörda i och med provens införande och eleverna uppskattade snabb feedback; de får nämligen resultatet direkt. Vissa initialsvårigheter med föråldrad datorpark och dålig

bandbredd fanns, men de är nu i huvudsak åtgärdade och skolorna anser i allmänhet att det tekniska fungerar smärtfritt. Relativt stora skillnader i resultaten mellan olika skolor och delar av landet har noterats, däremot inte mellan pojkar och flickor. I de enkäter som gjorts upplever dock fler pojkar än flickor att proven är lätta. Vissa elever tycker att textmassan är stor.

Testkonstruktörernas reflektioner är att

- det varken är lätt eller billigt att framställa datorbaserade prov
- det är en utmaning att få administratörerna i utbildningsväsendet, politikerna och testkonstruktörerna att samverka
- det är nödvändigt att ha ett nära samarbete med lärare, konstruktörer av datasystemen, statistiker osv.[5]
- skolorna är positiva till att delta i utprövningar
- det är en välsignelse att data genereras automatiskt (Moe, 2008; Norge).

Övningsmaterial finns tillgängligt på Internet och alla testtagare uppmanas att bekanta sig med detta för att undvika att ovana vid distributionssätt, uppgiftstyper, instruktioner etc. skall påverka resultatet (Norge).

Danmark har nationella prov, bl. a. engelska i åk 7. Proven är low-stakes för testtagaren och när utbyggnaden är klar kommer de att vara obligatoriska för skolorna att genomföra. Systemet kommer att finnas tillgängligt under en tidsperiod från februari till april och kapaciteten är ca 6000 samtidigt inloggade elever. Den enskilde läraren avgör när proven skall genomföras och det är inte nödvändigt att hela klassen gör provet samtidigt. Utöver detta erbjuds eleverna att göra två frivilliga prov under den övriga läsårstiden. Systemet kan på så vis mäta elevens utveckling.

Syftet är att ge information till elever och föräldrar om huruvida kursmålen har uppnåtts och att kartlägga i vad mån skolorna lyckas med sitt uppdrag. Det betonas dock att endast en del av elevens färdigheter mäts i dessa prov och att det är lärarens uppgift att kontinuerligt bedöma de övriga; t.ex. testas inte egen produktion i engelska. I alla prov prövas dock tre s.k. profilområden, vilket innebär att tre separata CAT-sessioner pågår samtidigt. Detta görs helt transparent för testtagaren som upplever att ett item slumpvis följs av nästa.

Ett system med renodlad, *item*-baserad CAT används. Intressant är dock att provet inte avbryts när systemet har gjort en statistiskt säkerställd bedömning, utan eleverna har 45 minuter på sig att göra så många uppgifter som de hinner. Denna tid kan förlängas för elever som eventuellt inte har hunnit göra nödvändigt antal items.

Starten skedde i maj/juni 2007 med reducerad item-bank. En revision hösten 2007 visade att item-banken inte var tillräckligt stor och att kvalitén på en hel del items för låg. Under 2008 utvecklades och utprövades nya items och inga prov genomfördes. Enligt planen skall de nya provuppgifterna pilottestas under våren 2009 och systemet skall vara färdigt och i drift under våren 2010. Full effekt anser man sig uppnå efter några år då man kan följa elevers och gruppers utveckling över tid på ett sätt som är unikt i världen.

Resultaten på alla nivåer (individ, grupp, skola osv.) är konfidentiella. Tanken med detta är

[5] Hittills har såväl provkonstruktörer som datateknisk personal funnits vid universitetet I Bergen, vilket har ansetts vara en stor fördel. Den tekniska sidan skall dock nu överföras till *Utdanningsdirektoratet* i Oslo.

delvis att undvika att undervisningen ändras så att man studerar för provet, jfr *washback* ovan. De som berörs av resultaten har dock rätt att se dem och det är alltså möjligt för t.ex. en kommun att göra jämförelser mellan skolor, rektor kan se den egna skolans alla resultat osv. Läraren får tillgång till en detaljerad rapport om varje elevs prestation, där det t.ex. framgår vilka items varje elev har fått och det är på så vis möjligt för lärare och elev att gå igenom uppgifterna tillsammans. Föräldrarna skall också informeras och datorn genererar för detta ändamål automatiskt en utförlig redogörelse (Danmark; Wandall, 2008). Kritik har riktats mot valet av ingenjörsfirman COWI som ansvariga för konstruktion av proven, samt mot de stora summor som lagts ner på dessa datorbaserade prov. Enligt kritikerna håller provuppgifterna inte måttet (Richter, 2009)

Island håller på att utveckla adaptiva prov. I första omgången (2009) kombineras P&P och CBA för åk 10 i ämnena matematik, isländska och engelska. År 2010 kommer adaptiva datorprov att genomföras i samma årskurs och om försöket faller väl ut, kommer även åk 4 och 7 att omfattas. Syftet är att informera elever, föräldrar och lärare om hur väl eleverna uppnår målen, samt att se hur skolor och nationen som helhet uppfyller ställda krav (Island).

CITO (*Nederländerna*) konstruerar nationella prov i bl.a. språk och matematik för olika stadier. Proven kan fås antingen som CBT eller P&P, till ca 90% används CBT. Tidigare distribuerades proven via Internet, men detta övergavs av säkerhetsskäl och numera tillhandahålls CBT-versionen på CD. Ett Cito-prov används inför gymnasievalet. Detta är dock omdebatterat och ca 20% av skolorna väljer alternativa test. Utveckling av nya prov för gymnasiets teoretiska linjer pågår; pilot 2010 (CITO).

USA: *Virginia Department of Education, Texas Education Agency* och *Idaho State Board of Education* har sina prov online med P&P som alternativ. Skoldistrikten avgör vilken distributionsmetod som skall användas. Det vanligaste skälet att man väljer P&P är begränsad tillgång på datorer. Språkproven är inte adaptiva.

Flera universitet och stora privata institut hemmahörande i USA har utvecklat adaptiva prov. De listas nedan under rubriken ”Prov utvecklade vid universitet och andra institutioner”.

Skottland Scottish Qualifications Authority (SQA) har inlett en brett upplagd satsning på *e-assessment* för såväl formativt som summativt inriktade syften (Skottland).

Nya Zeeland har ett slags nationella prov, asTTle . Dessa finns online från och med år 2008, tidigare levererades de på CD-ROM. Från hemsidan: ”It is an educational resource for assessing literacy and numeracy (in both English and Māori) developed for the Ministry of Education by the University of Auckland” (asTTle). En person som vi låter förbli anonym svarade så här på frågan vad lärarna på Nya Zeeland tycker om proven: ”Mixed feelings. It's a useful assessment that gives you norm-referenced information, and the tests can be generated quite quickly and to your own teaching/learning needs, but many schools seem to be using it more as assessment for collecting evidence of achievement rather than assessment to identify further learning needs.”

Kanada (*Alberta*, delstatsprov). I ett mycket ambitiöst projekt flyttades provverksamheten i Alberta från P&P till CAA (Computer Adaptive Assessment). Ett privat företag, Castle Rock Research, fick överta provbanken (Castle Rock). Resultatet har blivit ett ramaskri från lärare och lärarfack, som anser att kostnaden blir för hög och att de MCT-uppgifter som distribueras motverkar läroplanens tankar om att utveckla kreativitet. Vidare kritiseras att lärarfacken inte

fick vara med då besluten fattades. Av 150 000 elever som skulle ha omfattats av proven hade ca 4000 genomfört dem i februari 2007. De tidigare P&P-proven däremot hade stor acceptans (Teachers Alberta, 2007).

Prov utvecklade vid universitet och andra institutioner

GRE, the *Graduate Record Examination*. testar engelska och matematik och utvecklas av ETS, Educational Testing Service (GRE). ETS administrerar flera prov, bl. a. TOEFL, SAT och Criterion, en webbaserad skrivtjänst baserad på e-rater (se ovan). ETS är en av de stora aktörerna inom dator-baserade prov och mycket forskning är knuten till organisationen (GRE)

*TOEFL "*Test of English as a Foreign Language". Provet är inträdeskrav för många engelsk-talande universitet, främst i USA. Testet levereras antingen som P&P eller datorbaserat och distributionen av detta sker över säker internetförbindelse. Ett gratis linjärt exempelprov finns online (TOEFL). Proven var tidigare CAT, men vid en grundlig översyn 2005/2006 ändrades detta och de är nu linjära. De fyra basfärdigheterna testas med hjälp av dator. Ingen speciell instruktion ges för hanteringen av dator eller programvara. Färdigheterna *tala* och *skriva* bedöms efter senaste revisionen inte maskinellt. Testtagaren tillåts numera föra anteckningar under provets gång.

Anledningen till omarbetningen som gjordes var att validiteten ifrågasattes. Provet mäter nu färdigheter mer integrerat och ska bättre återspegla akademiska språkkrav. Syftet är att testa det språk man behöver som student i USA "från klassrummet till bokhandeln". Förmågan att kombinera och sammanställa information från olika källor är viktig; t.ex. testas *läsa* och *höra* genom en skriftlig eller muntlig sammanställning (TOEFL).

GMAT, The Graduate Management Admission Test, är ett datoradaptivt prov för att bedöma förutsättningarna för att lyckas med ekonomiska studier i USA. Det påminner bitvis om hög-skoleprovet i Sverige. Det finns t.ex. en del som verkar motsvara NOG. Provet mäter (med *multiple choice*-uppgifter) grundläggande verbal och logisk förmåga. En del består av analytiskt skrivande vilket först bedöms av maskin (ett program som kallas Intellimetric, se ovan), sedan av mänskliga bedömare (GMAT).

ACT/Compass College placement test. Ett test som ska hjälpa colleges och studenter att hitta styrkor och kompetensområden som behöver speciell uppmärksamhet och ge hjälp vid kursval, alltså en typ av formativt/diagnostiskt material. Beträffande ESL (*English as a Second Language*) testas grammatik/språkfärdighet, läsa och höra. Provet använder sig av *testlets* (se ovan; Chapelle & Douglas, 2006; ACT)

MELAB (Michigan English Language Assessment Battery) är ett språkligt inträdesprov till universitet i USA och Kanada som alternativ till TOEFL. (MELAB)

WebLAS, Language Assessment System, University of California har prov knutna till undervisning i engelska som andraspråk och främmande språk (japanska och koreanska) på universitetet. Flera institutioner samarbetar för att utveckla prov och bedriva forskning kring datorbaserade prov. Målsättningen är att en hög grad av autenticitet skall uppnås. Ett intres-sant inslag i proven är att man använder video för hörförståelse. En undervisningssituation med bl. a. whiteboard och PowerPoint-bilder visas. Proven är så flexibelt utformade att de lätt ska kunna anpassas till andra språk (WebLAS; Chapelle & Douglas, 2006).

IELTS är ett samarbete, påbörjat 1989, mellan University of Cambridge ESOL Examinations,

the British Council och IDP Education Australia. Två varianter finns, en som testar akademisk nivå för dem som vill studera på universitet, och en mer allmän för icke-akademiska, yrkesutbildningsrelaterade behov (IELTS).

BULATS, Business Language Testing Service. Gör språkprov för yrkeslivet. Resultatet av ett samarbete mellan University of Cambridge ESOL Examinations, Alliance Française, Goethe-Institut och Universidad de Salamanca (BULATS).

Cambridge, ESOL har provcentra över hela världen för olika certifikat: *Key English Test* (*KET*), *Preliminary English Test* (*PET*), *Business English Certificate* (*BEC*), *Preliminary level, Business English Certificate* (*BEC*), *Vantage level, Teaching Knowledge Test* (Cambridge ESOL).

Prov utvecklade av organisationer knutna till EU

Dialang-projektet är ett europeiskt projekt, finansierat av EU under det s.k. Sokrates-programmet, med syfte att ta fram diagnostiska språkprov i 14 olika europeiska språk. Öppet för alla och ett nedladdningsbart gratisprogram finns, dock endast för Windows. Många intressanta exempel på provtyper finns på hemsidan (Dialang).

Surveylang. ”European Survey on Language Competences”. Surveylang är ett projekt, där 32 länder är delaktiga. På EU:s uppdrag skall man testa fem språk, tre färdigheter (läsa, höra, skriva) och på sikt även tala. Provet kommer att vara datorbaserat för dem som vill, men P&P kommer att finnas som alternativ. Ursprungligen var tanken att proven enbart skulle vara datorbaserade, men det visade sig att flera länder helt enkelt inte hade den IT-infrastruktur som behövdes. Ca 1500 elever/land/språk ska genomgå provet.

Tidsplan: under 2009 skall mjukvaruplattformen utvecklas och utprövningar görs under 2010 År 2011 skall testning genomföras i hela Europa och slutrapporten är planerad till 2012. Tanken är att systemet ska rulla på den hårdvara skolorna har och det utvecklas som open-source-projekt, dvs. det är fritt att använda för all icke-kommersiell verksamhet. Alla allmänna instruktioner för provets genomförande kommer att vara på testtagarnas modersmål, medan instruktioner kopplade direkt till provuppgifterna blir på målspråket. Klientdatorerna ska startas (bootas) från en usb-sticka som har ett Linux-system. Då denna metod ger unika möjligheter att styra vad som kan/får göras i systemet är det väl värt att följa detta försök (Surveylang; Ryssevik, 2008; Bjerkestrand, 2008).

Övrigt

Alliance Française, *Goethe-institutet* och *Universidad de Salamanca* har färdighetsprov i respektive språk. Dessa distribueras av ett antal aktörer på marknaden.

CARLA, Center for Advanced Research on Language Acquisition, University of Minnesota, är finansierat av USA:s utbildningsdepartement och arbetar bl. a. för att förbättra språkundervisning och språkinlärning. Som ett led i detta utvärderar man kvalitén i existerande amerikanska *proficiency*-test i franska, tyska och spanska. CAT-kompetens finns. (CARLA)

Eurocentres ger undervisning och administrerar flera av de stora proven. Det finns många referenser till ett adaptivt ”Vocabulary Size Test” från Eurocentres, men bara andrahandsuppgifter har hittats.

Avslutning

"Technology is actually changing the way language is used and therefore the abilities required to use it" (Chapelle & Douglas, 2006)

Datorbaserad testning har många fördelar; multimediastöd vid prövning av hörförståelse, simuleringsuppgifter med autentiskt innehåll, pappersfri, elektronisk distribution, snabb och konsekvent bedömning osv. men vägen kantas av frågetecken vid nästan varje steg. Mäter datorbaserade prov färdigheter på ett tillförlitligt sätt och skapar de orättvisor mellan könen eller personer med olika kulturell bakgrund? Spelar datorvana någon roll? Är förutsättningarna olika för testtagare vid olika arbetsstationer? Hur är det med säkerheten vid high-stakes-prov? Kommer bedömningen av språkliga färdigheter i praktiken att reduceras till mer slutna uppgifter eller rena flervalsuppgifter?

Vad förlorar vi om vi väljer den enkla utvägen, nämligen att bygga ett prov på de traditionellt vanligaste uppgiftstyperna vid datorbaserad bedömning, d v s *MCT* och *matching* och avstår från andra, kanske bättre sätt att bedöma språkliga kompetenser? Som sades i inledningen lurar den ständiga fällan runt hörnet vid en övergång till datorbaserade prov, nämligen att mäta det lätt mätbara och den viktigaste frågan att ställa sig måste vara om datorn lämpar sig för att bedöma det vi vill bedöma och om det finns en risk att datorbedömning kan missa viktiga aspekter av den *construct* vi har definierat och om det finns andra, mer lämpliga sätt att bedöma dessa aspekter. Chapelle & Douglas (2006) skriver "… constraints placed on computer-adaptive testing can prescribe test tasks, making them poor measures of textual competence". Deras grundtes är att syftet med en övergång till datorbaserade prov bör vara att på ett bättre och mer autentiskt sätt bedöma språklig färdighet. "The authenticity analysis requires the researcher to examine the types of tasks and uses of language that examinees will be engaged in beyond the test setting" (s. 94).

Bland annat beroende på brist på pengar och expertis är vi ännu inte framme vid målet att skapa system och utvecklingsverktyg som är idealiska för språkbedömning: "… ideal authoring systems have not been developed for language assessments, but this is an active area of enquiry" (Chapelle & Douglas, 2006, s. 106) och vidare konstateras att trots den utveckling som skett av språktestning med hjälp av teknologi, kan inga revolutionerande förändringar av bedömning rapporteras, men möjligen kan det i framtiden ske en utveckling som är gynnsam.

Följande citat rör egentligen högre utbildning, men torde vara tämligen allmängiltigt:

'If lower-order learning is an unintended educational consequence of on-line assessment, then any perceived or real gains made in efficiency, staff workload reduction and/or cost savings may be counterbalanced by a significant drop in the quality of higher education outcomes.' (Assessing Learning in Australian Universities, 2002)

Källförteckning

ACT http://www.act.org/compass (Hämtad 28/10, 2008)

Alderson, J.C. (1990). Learner-centred testing through computers: Institutional issues in individual assessment. In J. De Jong & D.K. Stevenson (eds), *Individualizing the assessment of language abilities* (pp. 20-27). Clevedon, UK: Multilingual Matters.

Alderson, J.C. & Huhta, A. (2005) The development of a suite of computer-based diagnostic tests based on the Common European Framework, *Language Testing* 22(3), 301-320.

Alderson, J. C. & Wall, D., (1993). *Does washback exist*? *Applied Linguistics, 14,* 115-129.

Apple Inc. http://www.apple.com (Hämtad 28/10, 2008)

@ventures http://www.ventures.dk (Hämtad 28/10, 2008)

@venx http://uddannelsesforum2006.emu.dk/udstillere/oversigt/aventures.html (Hämtad 21/10, 2008)

Assessing Learning in Australian Universities: Ideas, strategies and resources for quality in student assessment (2002) http://www.cshe.unimelb.edu.au/assessinglearning/03/online.html (Hämtad 29/10, 2008)

asTTle http://www.tki.org.nz/r/asttle/ (Hämtad 27/10, 2008)

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly 2*(1), 1-34.

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

Bjerkestrand, Ø. (2008) http://crell.jrc.it/Presentations_Iceland%202008/Bjerkestrand.pdf (Hämtad 27/10, 2008)

Björnsson, J.K. (2008) The PISA Computer Based Assessment of Science: What Did We Learn? http://crell.jrc.ec.europa.eu/Presentations_Iceland%202008/Björnsson%20-%20CBAS.pdf (Hämtad 27/10, 2008)

Blackboard http://www.blackboard.com (Hämtad 21/10, 2008)

Bridgeman, B. (2008) Experiences from Large-Scale Computer-Based Testing in the USA, http://crell.jrc.it/Presentations_Iceland%202008/Bridgeman.pdf (Hämtad 30/10, 2008)

Brown, A. & Iwashita, N. (1996) Language background and item difficulty: the development of a computer-adaptive test of Japanese. Pergamon *System,* Vol. 24, No. 2, 199-206

Buck, G. (2001). *Assessing Listening.* Cambridge: Cambridge University Press.

BULATS http://www.bulats.org (Hämtad 29/10, 2008)

Bunderson, C. V., Inouye, D. K., Olsen, J. B. (1988) Abstract: The Four Generations of Computerized Educational Measurement, Report Number: RR-88-35 http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c3921509/?vgnextoid=7b13457727df4010VgnVCM10000022f95190RCRD&vgnextchannel=dcb3be3a864f4010VgnVCM10000022f95190RCRD (Hämtad 21/10, 2008)

Burstein, J. C., Kaplan, R. M., Rohen-Wolff, S., Zuckerman, D. I., Lu, C. (2000) A Review of Computer-Based Speech Technology for TOEFL 2000, ETS, Princeton, NJ RM-99-5 http://www.ets.org/Media/Research/pdf/rm-99-05.pdf (Hämtad 12/4, 2008)

Cambridge ESOL http://www.cambridgeesol.org/exams/exams-info/computer-based-testing.html (Hämtad 30/10, 2008), om item-banken: http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN23p3-5.pdf (Hämtad 30/10)

CARLA http://www.carla.umn.edu/about/lrc (Hämtad 29/10, 2008)

Castle Rock http://www.castlerockresearch.com/caa/, http://www.castlerockresearch.com/caa/FAQ.aspx (Hämtad 27/10, 2008)

CatGlobal http://www.catglobal.com/CATGlobal8/pdf/support/catswsys_8_2_rnotes.pdf (Hämtad 20/10, 2008)

Common European Framework of reference for languages: learning, teaching, assessment (2001) http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp (Hämtad 31/10, 2008)

Chapelle, C. A. & Douglas, D. (2006), Assessing Language through Computer Technology, Cambridge University Press

Chun, W. C. (2006) An Analysis of a Language Test for Employment: The Authenticity of the PhonePass Test, Language Assessment Quarterly, Volume 3, Issue 3 July 2006 , s. 295 – 306

CITO http://www.cito.com (Hämtad 29/10, 2008)

CitoTester http://www.cito.nl/com_comprod/citotester/eind_fr.htm (Hämtad 27/10, 2008)

Clariana, R. & Wallace, P. (2002) Paper-based versus computer-based assessment: key factors associated with the test mode effect, *British Journal of Educational Technology* Vol 33 No5, 593-602

Coniam, D. (1998) Voice recognition software accuracy with second language speakers of English. Faculty of Education, The Chinese University of Hong Kong, Sha Tin, Hong Kong. Copyright © 1999 Elsevier Science Ltd. http://eric.ed.gov:80/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=EJ589710&ERICExtSearch_SearchType_0=no&accno=EJ589710 (Hämtad 29/11, 2008)

COPI http://www.cal.org/topics/ta/copi.html (Hämtad 18/10, 2008)

Danmark http://evaluering.uvm.dk/templates/velkomst_layout.jsf (Hämtad 31/10, 2008)

Davis, A. (2003) Three heresies of language testing research.
http://ltj.sagepub.com/cgi/reprint/20/4/355 (Hämtad 13/12, 2008)

Dialang http://www.dialang.org/swedish/index.htm,
http://dialang.org/project/english/index.html (Hämtad 25/10, 2008)

Dialang *insertion*
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item7/item7.htm
(Hämtad 18/10, 2008)

Dialang indirekt tal med ljudalternativ
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item11/item11.htm
(Hämtad 18/10, 2008)

Dialang *Benchmarking* i färdigheten "skriva"
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item12/item12.htm
(Hämtad 18/10 2008)

Dialang *Benchmarking* i färdigheten "tala"
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item13/item13.htm
(Hämtad 18/10 2008)

Dialang *Confidence in chosen response* (slags självbedömning)
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item8/item8.htm
(Hämtad 20/10, 2008)

Dialang *Deletion*
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item8/item8a.htm
(Hämtad 20/10, 2008)

Dialang *Hot spot* Markera mening i textmassa samt lexikon
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item6/item6.htm
(Hämtad 20/10, 2008)

Dialang *Re-organisation*
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item5/item5.htm
(Hämtad 20/10, 2008)

Dialang Tematisk gruppering
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item9/item9.htm
(Hämtad 20/10, 2008)

Dialang Videoclip i hörförståelse
http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/item3/item3.htm
(Hämtad 20/10, 2008)

e-rater April 2005 issue of R&D Connections
http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=ed562d3631df4010VgnVCM10000022f95190RCRD&vgnextchannel=a330253b164f4010VgnVCM10000022f95190RCRD (Hämtad 17/10, 2008)

Enlight http://www.teststation.com/main/about_enlight (Hämtad 29/10, 2008)

Erickson, G. (2009). *Good Practice in Language Testing and Assessment – A Matter of Responsibility and Respect.* Paper presenterat vid the LTTC International Conference on English Language Teaching and Testing, Taipei, Taiwan.

Erickson, G. (2009; under publicering). ”Att bäras åt” – Om den goda bedömningens flerfaldighet och ömsesidighet. I Tornberg, U. m.fl. *Språkdidaktiska perspektiv. Om lärande och undervisning i främmande språk.* Liber (planerad utg. sept. 2009).

FastTEST pro http://www.assess.com/xcart/product.php?productid=273 (Hämtad 21/10, 2008)

GMAT http://www.gmac.com/gmac/thegmat/ (Hämtad 30/10, 2008)

GRE http://gre-exams.com/gre-tests-mons.html?gclid=CNjHhYGbzJYCFQ2Y1QodZSPmyg (Hämtad 29/10, 2008)

Halldórsson, A. M., McKelvie, P. & Björnsson, J. (2009). Are Icelandic Boys really better on Computerized Tests than Conventional ones? Interaction between Gender, Test Modality and Test Performance. I F. Scheuermann & J. Björnsson, (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (s. 201-208). European Commission Joint Research Centre. Luxembourg: Office for Official Publications of the European Communities.

Hot Potatoes http://web.uvic.ca/hrd/halfbaked/#latest (Hämtad 29/10, 2008)

IELTS http://www.ielts.org (Hämtad 29/10, 2008)

Intellimetric (2008) http://www.vantagelearning.com/school/products/intellimetric (Hämtad 30/10, 2008)

Island http://crell.jrc.it/Presentations_Iceland%202008/Björnsson%20-%20proposal.pdf http://crell.jrc.it/Presentations_Iceland%202008/Skulason.pdf (Hämtad 28/10, 2008)

Kane, M. (2006). Validation. I R. L. Brennan (ed.), *Educational Measurement* (Fourth edition, s. 17-64). Westport CT: American Council on Education/Praeger Publishers.

Kyllonen, P.C. (September, 2008). New Constructs, Methods, & Directions for Computer-Based Assessment. In The Transition to Computer-Based Assessment: Lessons Learned from the PISA 2006 Computer-Based Assessment of Science (CBAS) and Implications for Large-Scale Testing http://crell.jrc.ec.europa.eu/Presentations_Iceland%202008/Kyllonen.pdf (Hämtad 29/11, 2008)

Lee, Y.-W., Gentile C., Kantor, R. (2008) Analytic Scoring of TOEFL® CBT Essays:

Scores From Humans and E-rater, ETS, Princeton, NJ, 2008

Lee, Y.-W. (2007) (with M. Chodorow and Claudia Gentile) Seoul National University, Seoul, Korea Abstract presenterad vid TESL/Applied Linguistics: Conference on Technology for Second Language Learning, Friday, September 21 and Saturday, September 22, 2007 http://www.public.iastate.edu/~apling/TSLL/5th_2007/2007_abstracts2.html (Hämtad 20/10, 2008)

Lilley, M. (2007) The Development and Application of Computer-Adaptive Testing in a Higher Education Environment, The University of Hertfordshire (s. 79-91) http://linkinghub.elsevier.com/retrieve/pii/S0360131503001465 (Hämtad 16/5, 2008)

Linacre, J. M. (2000) Computer-Adaptive Testing: A Methodology Whose Time Has Come, MESA Psychometric Laboratory, University of Chicago http://www.rasch.org/memo69.pdf (Hämtad 20/9, 2008)

Longman Market Leader http://www.market-leader.net/ (Hämtad 23/10, 2008)

Macspeech http://www.macspeech.com/product_info.php?products_id=592

Magal-Royo, T. (2007) (University Polythecnic of Valencia, Spain) Teaching English as a Second or Foreign Language March 2007 Volume 10, Number 4 (Recension av Chapelle & Douglas)

Martin, R. & Binkley, M. (2009). Gender differences in cognitive tests: a consequence of gender dependent preferences for specific information presentation formats. I F. Scheuermann & J. Björnsson, (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (s. 201-208). European Commission Joint Research Centre. Luxembourg: Office for Official Publications of the European Communities.

MELAB http://www.lsa.umich.edu/eli/testing/melab (Hämtad 29/10, 2008)

Messick, S. A. (1989). Validity. I R. L. Linn (Ed.), *Educational Measurement* (s. 13-103). New York: Macmillian.

Moe, E. (2008) http://crell.jrc.it/Presentations_Iceland%202008/Moe.pdf (Hämtad 29/10, 2008)

Moe E., Carlsen C., & Hasselgren A. (2006). Digitale leseprøver i engelsk– utfordringer og muligheter. Norsk Pedagogisk Tidsskrift 5(90), 391-403

Norge, Norska nationella prov http://udir.no/templates/udir/TM_Tema.aspx?id=3418 (Hämtad 27/10, 2008)

NWEA http://www.nwea.org/about/index.asp (Hämtad 20/3, 2009)

Ockey, G. J., (2007) Construct implications of including still image or video in computer-based listening tests. 2007; 24; 517 *Language Testing*. DOI: 10.1177/0265532207080771. Online-version: http://ltj.sagepub.com/cgi/reprint/24/4/517 (Hämtad 24/11, 2008)

PARES http://www2.computer.org/portal/web/csdl/doi/10.1109/ICALT.2004.1357652 (Hämtad 29/10, 2008)

Pearson VUE http://www.pearsonvue.com/sponsors/dev/ (Hämtad 21/10, 2008)

POET/OEPT Demo och information finns på http://web.ics.purdue.edu/~aginther (Hämtad 29/10, 2008)

QuestionMark http://www.questionmark.com/us/perception/index.aspx (Hämtad 29/10, 2008)

Quiz Center http://school.discoveryeducation.com/quizcenter/quizcenter.html (Hämtad 29/10, 2008)

Renet Oy http://www.renet.fi/ (Hämtad 18/10, 2008)

Renstar http://www.renlearn.com/sr/ (Hämtad 6/2, 2009)

Respondus http://www.respondus.com (Hämtad 21/10, 2008)

Richter, L. (2009) Eksperter: Nationale test er spild af penge, Dagbladet Information, 6/4, 2009 http://www.information.dk/187374 (Hämtad 24/4, 2009)

Ripley, M. (2008) Gender related performance of 9 and 13 year-olds in mathematics and problem solving: computer-and paper-based tests http://crell.jrc.it/Presentations_Iceland 2008/Ripley - Gender.pdf (Hämtad 26/10, 2008)

Ryssevik, J. (2008) http://crell.jrc.it/Presentations_Iceland%202008/Ryssevik.pdf (Hämtad 27/10, 2008)

Scriven, M. (1967). The methodology of evaluation. In R. V. Tyler (ed.), *Perspectives of Curriculum Evaluation*, pp. 39-83. Chicago: Rand McNally.

Sibberns, H. (2008) Experiences with moving to computer-based testing. Test preparation and field operations in two computer-based assessment tests, IEA Data Processing and Research Center Reykjavik, 29.09.08

Skottland www.sqa.org.uk/e-assessment (Hämtad 5/5, 2009)

SpeechRater http://toefl.startpractice.com/programs/toefl/toefl_faq.htm (Hämtad 30/10, 2008) http://www.mkzechner.net/SLaTE07.pdf (Hämtad 30/10, 2008)

starQuiz Cosmic Soft http://cosmicsoft.net/starQuiz/index.html (Hämtad 27/3, 2009)

Surveylang http://www.surveylang.org/en/index.html (Hämtad 27/10, 2008)

Sørensen, H. & Møller Andersen A. (2009). How did Danish Students solve the PISA CBAS items? Right and Wrong Answers from a Gender Perspective. I F. Scheuermann & J. Björnsson, (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (s. 201-208). European Commission Joint Research Centre. Luxembourg: Office for Official Publications of the European

Communities.

Teachers Alberta (2007)
http://www.teachers.ab.ca/Issues%20In%20Education/Real%20Learning%20First/RLF%20Library/Pages/Computer%20Adaptive%20Assessment%20FAQs.aspx (Hämtad 27/10, 2008)

TOEFL
http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD (Hämtad 30/10, 2008);
http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf (Hämtad 27/2, 2009)

Tonidandel, S. & Quiñones, M. A. (2002) Reactions to Adaptive Testing: Effects of Test Length and Explanation A portion of this paper was presented at the annual meeting for the Society of Industrial/Organizational Psychologists, Toronto, Canada, April 2002

Wall, D. & Horák, T. (2006) The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 1, The Baseline Study Dianne Wall and Tania Horák, s.119 http://www.ets.org/vgn-ext-templating/v/?vgnextoid=8806c02084b1c010VgnVCM10000022f95190RCRD&vgnextchannel=d35ed898c84f4010VgnVCM10000022f95190RCRD The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe (Hämtad 29/11, 2008)

Wall, D. (2000) The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? Copyright © 2000 Elsevier Science Ltd. All rights reserved.

Wandall, J. (2008) http://crell.jrc.it/Presentations_Iceland%202008/Wandall.pdf (Hämtad 27/10, 2008)

Way, W. D., Davis, L. L., Fitzpatrick, S. (2006) Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments, Pearson rr0503
http://www.pearsonedmeasurement.com/downloads/research/RR_05_03.pdf (Hämtad 2/4, 2008)

WebLAS UCLA, WebLAS demonstration av hf med multimedia
http://www.humnet.ucla.edu/web/departments/alt/weblas_esl_demo/demo_listen_psych1_vid.htm (Hämtad 20/10, 2008), öppna svar: http://www.weblas.ucla.edu/ för exempel (klicka på am. knappen) (Hämtad 20/10, 2008)

Versant (PhonePass) http://www.ordinate.com/versant/versant.jsp (Hämtad 18/10, 2008)

Wikström, C. (2005) DATORBASERADE PROV - egenskaper, möjligheter och begränsningar, Umeå Universitet, BVM nr 14, 2005, ISSN 1652-7313

Wilhelm, O. & Schroeders, U. (2008) Traditional and Computerized Ability Measurement: Stressing Equivalence vs. Exploiting Opportunities The Transition to Computer-Based Assessment Reykjavik, Iceland 30.09.2008
http://crell.jrc.ec.europa.eu/Presentations_Iceland%202008/Wilhelm.pdf (Hämtad 20/10, 2008)

Bilagor

Programvaror för konstruktion och distribution

Viktiga parametrar vid ett eventuellt val av system bör vara hur flexibla lösningarna är; om ändringar i funktioner är möjliga och hur det kan utvecklas i framtiden för nya testtyper. Hanteringen av skrivtecken som inte ingår i den grundläggande uppsättningen av ASCII-tecken, som t.ex. å, ä, ö, ß och ç är också ett viktigt hänsynstagande. I synnerhet bör man titta på detta om amerikanska system skulle vara aktuella.

Vad gäller påpekanden om stöd för olika operativsystem nedan kanske man bör nämna att detta endast är relevant om en lösning med installation på alla testdatorer väljs, eller om man vill att konstruktion/distribution skall kunna ske från alla typer av datorer. Om en web-baserad lösning väljs, kan proven göras på alla plattformar, förutsatt att kodningen av sidorna sker enligt gällande standard och den som genomför kodningen är medveten om att detta är ett krav.

Här följer en förteckning över de programvaror och/eller företag vi har funnit som levererar lösningar för datorbaserade prov. Sannolikt finns det fler.

Hot Potatoes	Java-baserat[6], Windows, Mac, Linux. Detta är ett ganska enkelt program, huvudsakligen tänkt för lärare som vill skapa datorbaserade uppgifter (Hot Potatoes).
FastTEST pro	Detta är ett färdigt system för provkonstruktion och distribution. Det finns enbart för Windows. (FastTEST pro).
starQuiz	Et system från Cosmic Soft, framförallt tänkt för lärare som vill producera online-prov. Konstruktionsprogramvaran finns för både Windows och Mac OS. Systemet baseras på populära, öppna standarder. Proven kan laddas upp och publiceras på tillverkarens web-plats eller distribueras lokalt via server. Ett analysverktyg för elevernas resultat finns (starQuiz).
CitoTester	Administrationssystem för digitala prov. Enbart för Windows (CitoTester).
Respondus	Ett Windows-baserat verktyg för att skapa provuppgifter. Det kopplas till *e-learning*-system som ANGEL, Blackboard, eCollege, WebCT och IMS QTI. Företaget har också konstruerat en webbläsare (för Windows) som omöjliggör kopiering, utskrift och access till andra webbsidor (Respondus).
Quiz Center	QC är en gratis onlinetjänst där lärare efter att ha registrerat sig kan konstruera provuppgifter för elever (Quiz Center).

[6] Java är ett programmeringsspråk som fått mycket stor spridning och ofta används på Internet-sidor, men även i program som körs direkt i en dator. I sin rena form, ofta kallad SUN Java, är koden körbar på de flesta operativsystem, vilket förenklar utveckling och distribution. Nackdelen är att det kan vara relativt processorkrävande och därmed lätt kan upplevas som ”segt”.

PARES	PARES är ett grekiskt universitetsprojekt utvecklat i SUN Java[6], körd på PC, men det borde vara (eller gå att göra) plattformsoberoende, kräver dock en Oracle-databas (PARES).
QuestionMark™	Windows. Passar både för stora installationer och i mindre sammanhang (QuestionMark).
Blackboard	System som bl. a. innehåller testkonstruktionsverktyg. Finns endast för Windows (Blackboard).
TestStation™	Ett svenskutvecklat system från stockholmsbaserade Enlight AB. Körkortsprovet och datorkörkortet är två exempel på produkter från detta företag. Båda är CBT. Det är inte känt om det finns adaptiv kompetens (Enlight).
Pearson VUE	Tidigare Promissor. En stor aktör på marknaden med kompletta lösningar (Pearson VUE).
Renet Oy	Renet utvecklar datorbaserade prov, bl. a. *high-stakes* finska språkprov för invånare i Ingermanland som vill flytta till Finland. Svarstyperna sant/falskt och MCT maskinbedöms, för aktiva färdigheter används mänskliga bedömare. Ett prov i tyska för nivå B1 (CEFR[7]) har tagits fram. Det presenteras genom högupplöst video och ett tilltalande gränssnitt. Bedömningen av elevernas färdighet görs av mänskliga bedömare, enligt CEFR-skalorna. Vinsten från bedömningssynpunkt ligger i sparad tid, då bedömaren endast ser/hör svaren. En intressant iakttagelse är att testtagarna upplever det som mindre nervöst att ”prata med datorn” än att sitta ansikte mot ansikte med en bedömare. För närvarande finns ingen webbaserad lösning, utan programvaran måste installeras på varje dator där provet görs. En internetlösning är dock möjlig, om än på bekostnad av upplösningen I dagsläget endast för Windows, men lösningar för andra OS skulle kunna tas fram (Renet).
Dialang	Det CEFR-relaterade projektet[7] Dialang är det första kompletta systemet som syftar till diagnostisk bedömning. Tanken är att användare skall stimuleras till livslångt lärande och genom systemet få tillgång till stöd i denna process, bl. a. genom feedback på vad som kan förbättras och lämpliga områden för vidare studier. Det finns ett nedladdningsbart klientprogram (dock enbart för Windows och det kräver Internet-anslutning) och webbexempel på nya testtyper. Man kommer förhoppningsvis att utveckla provkonstruktionsverktyg (Alderson & Huhta, 2005; Dialang, 2008; se även CALT i världen nedan).

[7] Common European Framework of Reference for Language, ett dokument som bl.a. beskriver generella, för många länder gemensamma, referensnivåer vid språkbedömning. Språkfärdighet beskrivs i sex steg från A1(mycket basal nivå) till C2 (mycket hög nivå) (CEFR, 2001). Referensramen är översatt till mer än 30 språk, bl.a. till svenska.

CAT Software System	Computer Adaptive Technologies. Webbaserat eller *Client-server*-baserat (CatGlobal).
@venX	@ventures är underleverantörer till COWI av det danska nationella provet (@ventures).
STAR	Programsviten Star från Renaissance Learning används bl. a. i K12-prov i USA. Sviten omfattar Reading, Early Literacy och Math (Renstar).

Röstigenkänning:

Röstigenkänning innebär att datorn tolkar talat språk. Nedanstående program är inte nödvändigtvis kopplade till ett bedömningssystem. Det ”kompletta” systemet här är det under ”produktiva färdigheter” nämnda Versant.

Kurzweil Voice	Windows, (Coniam, 1998).
DragonDictate	Windows, (Coniam, 1998).
Dragon: NaturallySpeaking	Windows, (Wikström, 2005).
MacSpeech Dictate	Mac OS, (Macspeech).

Talsyntes:

Talsyntes innebär att datorn kan läsa upp skriven text. Systemen arbetar huvudsakligen med engelska och tekniken är kanske inte direkt relevant i nuläget, men skulle kunna bli en viktig del av framtida prov.

DECTalk	Windows
Mac OS X	Talsyntes av hög kvalitet inbyggd i systemet (endast engelska)