

Bedömningskompetens i AI:s tidevarv: Exemplet automatiskt genererade flervalsfrågor¹

Maria Rydell¹

¹Stockholms universitet

Abstract

Generativ AI och dess konsekvenser för utbildningssektorn har de senaste åren diskuterats flitigt. Föreliggande studie belyser automatiskt genererat språk och konsekvenser för lärares bedömningskompetens (*assessment literacy*, t.ex. Fulcher, 2012). Mer specifikt analyseras hur sex sfi-lärare och sex professionella provkonstruktörer värderar automatiskt genererade flervalsfrågor för att bedöma läsförståelse på sfi. Baserat på enkätsvar och uppföljande semi-strukturerade djupintervjuer diskuteras deltagarnas syn på de automatiskt genererade flervalsfrågorna när det gäller frågornas relevans, användbarhet och möjligheter och begränsningar ur ett pedagogiskt perspektiv. Även relationen mellan AI och mänskligt omdöme diskuteras. Resultaten visar att deltagarna såg pedagogiska möjligheter med de automatiskt genererade flervalsfrågorna men att de behöver granskas och justeras. För att kunna granska och värdera automatiskt genererade flervalsfrågor menar deltagarna att det krävs kunskap om provkonstruktion och andraspråksutveckling, samt erfarenhet av målgruppen. Flervalsfrågor är en vanligt förekommande testteknik för att bedöma läsförståelse. Den ökade förekomsten av digitala prov med automatiskt genererade delar kommer troligtvis att öka förekomsten av flervalsfrågor i standardiserade prov. Resultaten visar på hur ett format som är enkelt att rätta kan vara svårt att bedöma, när man med bedöma avser förmågan att kunna tolka och använda sig av resultaten. Avslutningsvis diskuteras hur förmågan att granska och värdera AI-genererat språk är en viktig del av lärares bedömningskompetens.

Inledning

AI och dess konsekvenser för utbildningssektorn har under de senaste åren diskuterats flitigt (se t.ex. Forsell, 2023; Voss m.fl., 2023; Winerö, 2022). Vilken roll automatiskt genererat språk i vid mening kommer att spela i framtiden är

oklart men något som lärare och andra som arbetar med språkbedömning behöver förhålla sig till redan nu. Mycket fokus har riktats mot minskad användning av skriftliga hemuppgifter och hur den nya tekniken kan användas på ett pedagogiskt sätt, till exempel genom automatiserad bedömning (t.ex. Skolverket, 2024; SOU 2022:17). Ännu finns lite forskning om kopplingen mellan generativ AI och bedömning, särskilt i en svensk kontext (Forsell, 2023). Den här artikeln undersöker vilka konsekvenser AI har för lärares bedömningskompetens (eng. *assessment literacy*). Mer specifikt studeras automatiskt genererade flervalsfrågor i syfte att bedöma läsförståelse i svenska som andraspråk och hur lärare och professionella provkonstruktörer resonerar om deras användbarhet, möjligheter och begränsningar. Eftersom användningen av flervalsfrågor i bedömning av läsförståelse bara kan vara relevant i förhållande till ett givet syfte och sammanhang har vi valt att rikta in oss mot den avslutande kursen i kommunal vuxenutbildning i svenska för invandrare (hädanefter används förkortningen sfi).

Sfi var tidigare en egen skolform men är sedan 2016 en integrerad del av komvux. Utbildningen riktar sig till deltagare över 16 år, skrivna i en kommun och i behov av att utveckla grundläggande färdigheter i svenska. Målgruppen är mycket heterogen när det gäller deltagarnas språkliga och utbildningsmässiga bakgrund. Utbildningen är därför indelad i tre studievägar beroende på deltagarnas vana av att studera och förväntade progressionstakt. Kurs D är den avslutande kursen inom samtliga studievägar. Syftet med sfi är att deltagarna ska utveckla en allsidig kommunikativ språkförmåga och kursplanen beskriver de språkliga färdigheter eleverna ska utveckla i muntlig produktion, muntlig interaktion, läsförståelse, hörförståelse och skriftlig färdighet (SKOLFS 2021:61). Till kurserna B, C och D finns nationella prov som innefattar delprov som prövar de olika färdigheterna. Flervalsfrågor är en vanligt förekommande testteknik för att bedöma receptiva färdigheter som läsförståelse och hörförståelse och förekommer både i övningsmaterial och i de nationella proven. Det är också ett svarsformat som lämpar sig för automatisk rättning. I och med en ökad förekomst av digitala prov² inom skolväsendet och ett önskemål om ökad andel automaträttade uppgifter är det ett svarsformat som därför troligtvis kommer att öka (Skolverket, 2020).

Den här studien är en del av projekt SWE-QUEST³ som dels syftar till att automatiskt generera flervalsfrågor på svenska, dels utvärdera frågornas användbarhet med tänkta användare, vilket i detta fall är lärare och provkonstruktörer. Föreliggande artikel är en del av projektets andra del som

handlar om utvärdering. I utvecklingen och utvärderingen av automatiskt genererade flervalsfrågor har forskning främst riktats in sig på de tekniska möjligheterna och begränsningarna (se t.ex. Kalpakchi & Boye, 2023; Kalpakchi & Boye, 2024). Mindre fokus har riktats mot utvärdering ur ett pedagogiskt perspektiv med tilltänkta användare.

De specifika frågeställningarna är följande:

- Hur värderar lärare och provkonstruktörer automatiskt genererade flervalsfrågors kvalitet ur ett bedömningsperspektiv?
- Vilka pedagogiska möjligheter och begränsningar ser lärare och provkonstruktörer med automatiskt genererade flervalsfrågor?
- Vilka konsekvenser har automatiskt genererade flervalsfrågor för lärares bedömningskompetens i ett vidare perspektiv?

Bedömningskompetens

Bedömning har en given plats i lärares yrkesvardag och kunskap om bedömning är en central del av lärares yrkeskunnande (Popham, 2009). Inom andraspråkundervisning, precis som i övriga skolväsendet, har bedömning fått en ökad roll. Språktestforskaren Glenn Fulcher (2012) pekar på tre orsaker till denna utveckling inom andraspråksundervisningen. Två faktorer menar han ligger utanför utbildningssektorn och en är driven inom utbildningssektorn. De ”yttre” faktorerna är en ökad redovisningskultur inom utbildningsväsendet med ökade krav på dokumentation och uppföljning, och en ökad förekomst av att använda språkkrav i migrationsprocessen. Inom utbildningssektorn har bedömningsroll för lärande gjort att bedömning fått en mer central roll i planering och genomförande av undervisning. Sammantaget innebär det att lärares bedömningskompetens (eng. *assessment literacy*) har fått en ökad roll.

Bedömningskompetens brukar definieras som både (praktisk) kunskap *i* bedömning och (teoretisk) kunskap *om* bedömning. Det handlar då om kunskap i (och om) provkonstruktion, bedömning av elevprestationer, tolkning och användning av bedömningsresultat, att kunna använda bedömningsresultat för att stödja pedagogiska beslut och stödja lärande. Begreppet innefattar även att kunna rapportera och kommunicera resultat, kunskap om lämpliga och olämpliga sätt att använda bedömning och testning, samt etiska perspektiv på testning och

bedömning (Akthar m.fl., 2021; Fulcher, 2012; Pastore, 2023). Till denna lista är det numera rimligt att lägga till kunskap om bedömning i relation till AI.

Meijer m.fl. (2023) menar att bedömningskompetens bör ses som ett dynamiskt koncept som formas i samspel med andra och utvecklas över tid och som är kopplat till lärares uppfattningar om sig själva och sitt uppdrag. I en föreslagen modell av Pastore och Andrade (2019) argumenterar författarna för att bedömningskompetens bör ses som ett multidimensionellt konstrukt bestående av vad de benämner a) *conceptual knowledge* – lärares uppfattningar om bedömning, lärande och undervisning, b) *praxeological dimension* – bedömning i praktiken och hur lärare integrerar bedömning i undervisning och i relation till undervisningsrelaterade mål och c) *socioemotional dimension* – lärares medvetenhet om vad deras position som bedömare innebär, bedömningens olika betydelse för elever och etiska aspekter av bedömning.

Nuvarande lättillgängliga generativa språkmodeller gör att det också finns ett behov av att diskutera en särskild "AI literacy" (Ng m.fl., 2021). I en forskningsöversikt framhålls att AI-literacy innefattar a) kunskap och förståelse, b) användning och tillämpning, c) utvärdering, d) skapandet eget material och e) etisk kunskap om användning (Ng m.fl., 2021). Som denna korta genomgång visat är bedömningskompetens ett mångfasetterat fenomen som också hänger ihop med att bedömning används för olika syften, i olika sammanhang och med olika typer av konsekvenser. Definitionen ovan av AI-literacy liknar definitionerna som används för bedömningskompetens. Förmågan att kritiskt granska och använda AI-genererat språk i bedömningssammanhang kan ses som en ny och viktig del i språklärares bedömningskompetens. Det faktum att generativa språkmodeller som ChatGPT3 och senare utvecklingar har blivit lättillgängliga gör att de kan ses som ytterligare en "yttre faktor" i Fulchers andemening som leder till ett ökat och till viss del nytt fokus på bedömningskompetens.

Bedömning av läsförståelse och flervalsfrågor

Dagens andraspråksundervisning bygger på en kommunikativt inriktad språksyn. Kommunikativ språkförmåga handlar om att kunna använda språket på ett anpassat sätt i olika sammanhang, för en mängd olika syften och mottagare (Rydell, 2021). Läsförståelse kan ses som en del av en mer övergripande

kommunikativ språkförmåga. I likhet med definitioner av kommunikativ språkförmåga handlar samtida definitioner av läsförståelse om att kunna läsa olika typer av texter, i olika typer av sammanhang och för olika syften. I andraspråksammanhang har avkodning, läshastighet och allmän språkförmåga i det givna språket betydelse. Särskilt ordförståelse har en central del i läsförståelse (Nation, 2001). För att kunna förstå en text behöver en stor del av det språkliga innehållet vara begripligt för andraspråksläsaren (Alderson, 2000).

Läsförståelse ses som ett komplext och sammansatt konstrukt (Alderson, 2000; Tengberg, 2018; Weir & Khalifa, 2008). Förutom att skilja på läsning av olika texttyper (t.ex. löpande och icke-löpande text) och olika syften med läsning (t.ex. sökläsning, läsa för att förstå huvudpoänger) är ett sätt att ringa in vad läsförståelse innebär att ta fasta på vad man gör med sin läsning och hur detta hänger ihop med hur information presenteras i texter. Ett ramverk som fått stor genomslagskraft är det som används i PISA (OECD, 2019). I PISA används begreppet *reading literacy*, vilket definieras som “understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one’s goals, to develop one’s knowledge and potential and to participate in society” (OECD, 2019, s. 28). Vidare specificerar ramverket både texttyper och olika kognitiva processer, så kallade läsprocesser. Dessa är a) *lokalisera information*, b) *förstå* (förstå uttryckt information samt även sammanföra information och dra slutsatser) och c) *värdera och reflektera*. Dessa olika läsprocesser antas innebära olika grad av komplexitet. I en studie där lärare fick identifiera och reflektera kring vilka läsprocesser som olika frågor i ett läsförståelseprov prövade framkom att lärarna inte var helt överens (Tengberg, 2018). Huruvida dessa kognitiva läsprocesser existerar som diskreta och urskiljbara enheter eller inte är alltså ingen självklarhet och de bör, menar Tengberg, användas med viss försiktighet. Tengberg (2015) argumenterar för att läsförmåga är ett vidare begrepp än läsförståelse och en bättre översättning av ”reading literacy”. I den här texten har jag valt att använda begreppet *läsförståelse* då det är vanligt förekommande inom just andraspråksbedömning och det begrepp som används i kursplanen i sfi. Det är också förståelse av texten som är i fokus för de automatiskt genererade flervalsfrågorna som används i denna studie.

Läsförståelse anses vara svårbedömt eftersom det är en osynlig process som sker i hjärnan. För att synliggöra läsförståelse behöver därför lärare och provkonstruktörer använda sig av olika tekniker. En vanligt förekommande testteknik är flervalsfrågor. En flervalsfråga består av en fråga, ett rätt svar och

ett antal distraktorer (trovärdiga men felaktiga svar). Det räknas som ett stängt format där svarsalternativen är på förhand givna.

Vid bedömning av språkförmåga, precis som vid all annan pedagogisk bedömning, är validitet centralt. Validitet kan enkelt uttryckt definieras som giltigheten och kvaliteten på de slutsatser som dras vid en bedömning eller ett resultat. Vid användandet av stängda svarsformat som flervalsfrågor är det lätt att få fram ett resultat som 27 rätt av 40. Men vad betyder ett sådant resultat? Bedömning handlar om att göra inferenser (McNamara & Roever, 2006). Särskilt i kommunikativ språktestning är man intresserad av att kunna dra slutsatser om elevers kommunikativa språkförmåga utanför själva testsituationen. Det innebär att bedömning alltid innefattar tolkning, en tolkning som användare behöver argumentera för. En av de mest välciterade definitionerna av validitet inom språkbedömning lyder att validitet är ”an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences and actions* based on test scores or other modes of assessment“ (Messick, 1989, s. 13, kursiv i original). Att ett läsförståelseprov mäter de olika kognitiva processer som definieras i ett konstrukt för läsförståelse hänger vidare ihop med det som kallas kognitiv validitet (Weir & Khalifa, 2008). Messick betonade även att validitetsbegreppet innefattar sociala värderingar och konsekvenser av bedömning (McNamara & Roever, 2006; Messick, 1989). Det kan exempelvis handla om hur bedömningar används, men också vilken syn på kunskap som bedömning förmedlar. Att analysera flervalsfrågor ur ett pedagogiskt perspektiv är således en komplex uppgift.

Fördelen med flervalsfrågor är att det är ett vanligt format internationellt och därmed ett format som många känner igen. Det kräver inte heller att testtagaren använder sig av produktiva färdigheter som att skriva, vilket kan vara en fördel av flera skäl. Det gör att man kan ha ganska många frågor, vilket är viktigt för mätsäkerheten (Wikström, 2013). Att elever inte behöver skriva svaren gör också att det blir mindre tolkning för den som rättar och att man har bättre möjlighet att ringa in själva läsförståelsen utan att elevers produktiva färdigheter ska påverka. På sfi är detta viktigt då just skriftlig färdighet är en svårighet för många och särskilt för elever med kort tidigare skolbakgrund (Institutionen för språkdidaktik, 2020). Välskrivna flervalsfrågor karaktäriseras av att de fångar relevant kunskap och är *diskreta*, det vill säga man ska kunna svara på dem var för sig och de ska inte vara beroende av varandra så att ett felaktigt svar gör att man inte kan svara på de nästkommande frågorna. Flervalsfrågor bör även vara

symmetriska när det gäller formen (som längd och grammatisk form) så att inte ett svarsalternativ sticker ut. Man försöker även skala bort redundansen så att inte samma information upprepas, vilket bidrar till att flervälsfrågorna kan bli så korta och tydliga som möjligt (t.ex. ALTE, 2005; Wikström, 2013). Negationer i frågeställningen bör undvikas. Vid provkonstruktion försöker man även undvika så kallad lexikal överlappning, det vill säga att exakt samma formulering förekommer i fråga/svarsalternativ och text för att undvika att rätt svar endast fås genom identifikation i texten snarare (såvida det inte handlar om informationssökning) än genom förståelse. Om lexikal överlappning används försöker man göra det medvetet och med symmetri så att alla svarsalternativ nämns. Ett alltför krångligt språkbruk i själva frågeformuleringarna kan också göra frågan onödigt svår, vilket kan innebära ett hot mot validiteten genom så kallad *konstruktirrelevant varians*, det vill säga att bedömningen påverkas av faktorer som ligger utanför konstruktet (Messick, 1989).

Att göra bra och välfungerande flervälsfrågor är tidskrävande och svårt (Alderson, 2000). Men när de väl är gjorda är det ett lättträttat format som också lämpar sig för automaträttning. Det finns dock även flera nackdelar med flervälsfrågor. Formatet möjliggör gissande och att man kan använda uteslutningsmetoden. Därför kan det bli till ett format som gynnar de som är ”test wise” och leda till att lärare fokuserar på att undervisa elever strategier för att ta sig an flervälsfrågor (Alderson, 2000), snarare än att fokusera på själva förståelsen. Det är också ett format som skiljer sig från läsning i icke-provsammanhang, vilket kan påverka vilka slutsatser som kan dras. Rupp m.fl. (2006) menar att genom sin konstruerade form påverkar flervälsfrågorna konstruktet och vad som bedöms, vilket i förlängningen kan påverka validiteten om inte bredden i konstruktet fångas. Distraktorerna kan också göra att testtagarna leds in på spår som de inte själva hade tänkt på (Alderson, 2000). En annan kritik som framförts är att frågorna kan tendera till att främst rikta in sig på enklare kognitiva processer (Wikström, 2013).

En flervälsfrågas svårighetsgrad är alltid ett samspel mellan fråga, svarsalternativ och texten. En viktig aspekt är hur själva relationen mellan flervälsfrågan och texten ser ut, om den är explicit eller implicit och om svaret finns i en del av texten eller om information mellan olika delar av texten behöver sammanföras samt hur distraktorerna förhåller sig till varandra och till texten (Kirsch & Mosenthal, 1995). Ytterligare en variabel är själva testtagaren och dennes förkunskaper (OECD, 2019). För provkonstruktörer innebär det en särskild

svårighet att skriva trovärdiga distraktorer som lockar de svagare läsarna men inte de starkare (Alderson, 2000). Utprövning av frågorna på målgruppen med efterföljande analys är därför viktigt för att kunna avgöra om en flervalsfråga fungerar och vilken svårighetsnivå den egentligen har.

Metod och material

Följande studie bygger i huvudsak på tolv semistrukturerade djupintervjuer med sex sfi-lärare och sex professionella provkonstruktörer. Intervjuerna genomfördes 2023-2024 och tog mellan 44 och 80 minuter. Lärarna hade många års erfarenhet av att undervisa på sfi eller svenska som andraspråk på komvux. Flera av deltagarna hade gått en kurs i provkonstruktion och bedömning och ansågs därmed ha fördjupad kunskap om ämnet. Provkonstruktörerna hade alla bakgrund som andraspråkslärare och arbetade med att göra nationella prov i sfi.

Underlaget för utvärderingen baseras på Kalpakchi & Boye (2023). Som grund för genereringen av de automatiskt genererade flervalsfrågorna låg 96 texter hämtade från frisläppta tidigare nationella prov i sfi, kurs D. Dessa texter valdes då de genomgått granskning i provkonstruktionsprocessen, vilket fick bli en garant för att texterna kan anses vara anpassade efter målgruppens nivå. Baserat på denna textkorpus genererades 717 flervalsfrågor mellan december 2022 och januari 2023 genom användandet av språkmodell GPT-3 (en sorts föregångare till den öppna ChatGPT, se Kalpakchi & Boye, 2023). Vid genereringen var 0,11 % av språkmodellens träningsmaterial text på svenska. Instruktionen som gavs var:

- fyra (4) svarsalternativ ska genereras
- bara ett (1) alternativ ska vara korrekt
- de andra tre alternativen ska vara inkorrekta men troliga
- frågan måste kunna besvaras utifrån texten

Metoden för generering och resultaten redovisas mer utförligt i Kalpakchi & Boye (2023). 44% av de genererade frågorna ansågs vara tillräckligt välfungerande utifrån kriterierna ovan. Värderingen gjordes dock inte av pedagoger med expertis i andraspråksbedömning varför det kan vara intressant att utvärdera frågorna med ett pedagogiskt fokus.

I föreliggande studie utvärderas några av frågorna av personer med stor vana av att bedöma läsförståelse på sfi. Baserat på den redovisade studien (Kalpakchi &

Boye, 2023) valdes fyra texter och tio automatiskt genererade flervalsfrågor ut. Det övergripande syftet med utvärderingen var att kunna fånga de automatiskt genererade flervalsfrågornas pedagogiska möjligheter och begränsningar. Dessa item (deluppgifter) valdes utifrån att de skulle pröva olika typ av läsning. Några frågor med vanligt förekommande problem som att fler svar kunde vara rätt inkluderades. Även mänskligt konstruerade flervalsfrågor kan ha problem och behöva revideras för att de ska bli välfungerande (Alderson, 2000). Flervalsfrågor som uppenbarligen inte fungerade valdes bort.

Inför intervjuerna fick deltagarna fylla i en enkät som de sedan fick tänka högt kring i uppföljningsintervjun. I enkäten användes de utvalda texterna och frågorna och deltagarna fick ta ställning till flervalsfrågornas relevans, svårighetsgrad, om endast ett svarsalternativ kunde anses vara rätt och distraktorernas trovärdighet. I anslutning till varje flervalsfråga fick deltagarna ta ställning till vad de tyckte att frågan prövade genom att svara på vad läsaren behöver kunna göra för att svara på frågan utifrån följande alternativ (det var möjligt för deltagarna att fylla i flera av svarsalternativen):

- lokalisera information som uttrycks explicit i texten
- matcha information i fråga, svarsalternativ och text som uttrycks med synonymer
- sammanföra information som uttrycks explicit i olika delar av texten
- dra slutsatser utifrån information som uttrycks i olika delar av texten
- annat (frisvar)

Intervjuerna utgick från ett intervjuschema indelat i tre tematiska delar. I den första delen fick deltagarna berätta om sina erfarenheter av att bedöma läsförståelse på sfi, resonera om möjligheter och utmaningar med att bedöma läsförståelse och hur de såg på för- och nackdelar med att använda flervalsfrågor för att bedöma läsförståelse. I den andra delen fick de tänka högt kring olika ställningstaganden i enkäten och resonera kring sina intryck av de automatiskt genererade flervalsfrågorna. I den tredje delen av intervjun fick deltagarna resonera kring möjligheter och utmaningar med automatiskt genererade flervalsfrågor och vilka pedagogiska konsekvenser de kunde se. Intervjuerna har transkriberats och har analyserats genom en innehållsanalys (t.ex. Bengtsson, 2016). Genom att ha lyssnat på intervjuerna vid upprepade tillfällen har en förtrogenhet skapats med materialet. Teman har skapats både deduktivt genom intervjuguiden och utifrån teoretiska perspektiv på bedömningskvaliteter, och induktivt genom de perspektiv som deltagarna själva lyfte fram. Nedan har både

de öppna svarsalternativen i enkäten och resonemangen i intervjuerna analyserats. Intervjuerna har genomförts med informerat samtycke. Deltagarnas namn har pseudonymiserats och benämns utifrån deras yrkesroll (Lärare och Provkonstruktör).

Resultat

Även om föreliggande studie är en kvalitativ studie och det inte går att dra några större slutsatser utifrån en enkät med tolv deltagare är det ändå intressant att konstatera att det var spridda uppfattningar när det gällde flervalsfrågornas användbarhet och vad de prövade. Nedan diskuteras deltagarnas uppfattningar om flervalsfrågornas kvalitet ur ett bedömningsperspektiv. Mer specifikt diskuteras flervalsfrågornas relevans, svårighetsgrad och trovärdighet samt vad de prövade. Därefter diskuteras uppfattningar kring de automatiskt genererade frågornas möjligheter och begränsningar ur ett pedagogiskt perspektiv samt relationen mellan generativ AI och mänskligt omdöme och konsekvenser för lärares bedömningskompetens.

Uppfattningar om flervalsfrågornas relevans, svårighetsgrad och trovärdighet

En övergripande tendens i enkäten var att frågan i flervalsfrågan ansågs vara relevant men att problem ofta uppstod i relation till texten och svarsalternativen. Som Provkonstruktör 2 uttrycker det i intervjun: ”Den klarar av att ställa en hyfsat relevant fråga men inte svarsalternativen”. Av de 10 automatiskt genererade flervalsfrågorna var det bara en fråga som fick en övervägande del nej-svar på frågan om frågan var relevant. Överlag var deltagarna överens. I två fall svarade alla 12 deltagare ja på om frågan kunde anses vara relevant. Om alla svar (10x12) läggs samman på frågan ”Är frågan relevant” blev resultatet 74,6 % JA och 25,9 % NEJ.

Frågan om uppfattad svårighet visade sig vara svår att svara på då en fråga kan vara svår av olika skäl, där vissa skäl kan ses som legitima (t.ex. att den kräver svårare processer som att sammanföra information och dra slutsatser) medan andra är icke-legitima ur provkonstruktionssynpunkt, som att frågan inte gick att svara på. Frågan om distraktorernas trovärdighet är dock intressantare. Att skriva trovärdiga distraktorer som fungerar, vilket ur ett provkonstruktionsperspektiv innebär att de ska vara felaktiga men tillräckligt trovärdiga i förhållande till texten

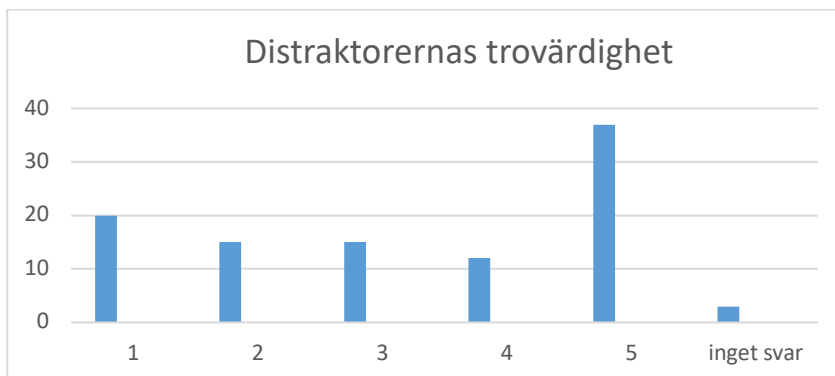
och därmed locka de svagare läsarna men inte de starkare, kan vara svårt även för professionella provkonstruktörer (Alderson, 2000). En tendens i enkäten var att uppfattningen om distraktorerna var trovärdiga skilde sig åt mellan deltagarna, även bland de professionella provkonstruktörerna. Enkätsvaren till en av flervalsfrågorna kan illustrera hur samma distraktorer värderades olika. 1 = låg grad och 5 = hög grad.

Figur 1. Distraktorernas trovärdighet som värderades kopplade till en flervalsfråga. Antal svar.



Detta mönster går igen när alla svar slås samman. I Figur 2 nedan illustreras den sammantagna bilden av distraktorernas övergripande grad av trovärdighet.

Figur 2. Distraktorernas trovärdighet – alla svar sammanslagna mätt i procent (10x12 svar). 1 = låg grad och 5 = hög grad.



Figur 2 visar att distraktorerna av flera deltagare och vid flera tillfällen ansågs trovärdiga, men också att uppfattningarna var spridda. Det var inte vid någon fråga som deltagarna var helt eniga om distraktorernas trovärdighet. Frågan om distraktorernas trovärdighet är central för om en flervälsfråga anses fungera eller inte. Om distraktorerna inte är trovärdiga kan det uppmuntra till att testtagarna använder sig av uteslutningsmetoden. När det gäller standardiserade prov avgörs frågan om ett item fungerar eller inte i regel efter en utprovning men det är också ett vanligt förfarande att använda sig av expertgruppers granskningar innan en provversion går ut på utprovning. Men i klassrumsbedömning görs i regel inga utprovningar. Lärare behöver därför kunna granska frågor för att avgöra om de fungerar och vad de prövar, vilket är en del av deras bedömningskompetens.

Uppfattningar om vad frågorna prövar

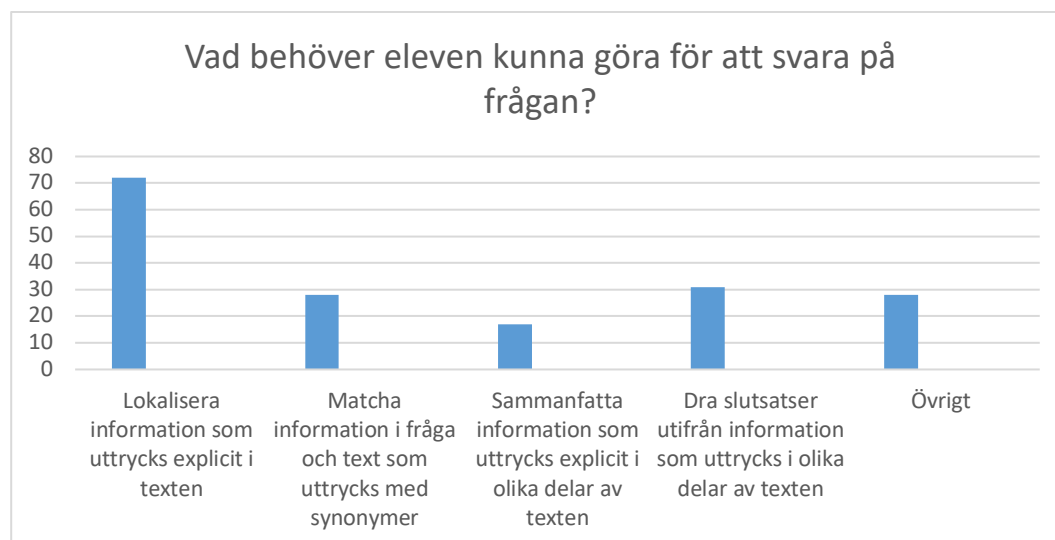
En återkommande uppfattning var att flervälsfrågorna, både frågekonstruktionerna och svarsalternativen, var för textnära. Lärare 6 menar att det ”mest är ytkunskap som testas” och att det är ”våldigt mycket explicit innehåll som efterfrågas”. Ett genomgående problem ansågs den lexikala överlappningen vara. Sammantaget leder det till, som majoriteten av deltagarna är inne på, att frågorna blir för lätta: ”man efterfråga synonymer, det finns nästan inte, det är explicit lexikal överlappning och det blir alldeles för enkelt om frågorna är konstruerade på det sättet.” (Lärare 6). De flesta flervälsfrågorna bedömdes i huvudsak pröva förmågan att lokalisera explicit information och många frågor ansågs därför vara för lätta: ”det var mitt första intryck /.../ det *är* lite lätt” och ”passar perfekt för kurs B” (Lärare 1). Lärare 1 resonerar vidare i intervjun kring sitt övergripande intryck på följande vis:

Ja jag tänkte att alltså om jag tänkte på just teknikerna och processerna som vi skulle ha aspekterna lokalisera, matcha, sammanfatta dra slutsatser /.../ dom första frågorna hade just det att du ska gå och kolla i texten så om vi tänker på just kunskapskravet för läsa D att hitta information är inte det mest viktigaste där så vi kräver också från eleverna i klassen att dom skulle tänka lite grann och då tänkte jag men det här är ju information som är superlätt att hitta /.../ även om en elev som inte har så mycket förståelse på nivå D, även om det inte är en elev som inte skulle förstå så mycket texten /.../ då är det lätt att hitta informationen. (Lärare 1)

Att kunna svara rätt på förståelsefrågor utan att förstå så mycket av texten kan innebära ett hot mot validiteten. Den sammantagna (10 x 12 svar) bilden av

deltagarnas syn på vad frågorna prövar visar på en övervikt för det mer textnära alternativet ”lokalisera information som uttrycks explicit i texten”.

Figur 3. Sammanställning över uppfattning om vad frågorna prövar mätt i procent.



I kategorin ”Övrigt” hamnade fritextalternativ som exempelvis ”hitta samma ord (overlap)”, ”räkna”, ”eleven kan inte svara på frågan” och ”gissa”.

Frågor som syftade till att pröva andra aspekter av läsförståelse som att dra slutsatser uppskattades. Dessa frågor var också de frågor som väckte mest diskussion. Som exempel ska vi nu djupdyka i en av de frågor som flera av deltagarna lyfte i intervjuerna (se Figur 4 nedan).

Figur 4. Exempeluppgift – automatiskt genererad flervalsfråga.

Hej,

Min mobil verkar äntligen fungera nu när jag har betalat fakturapåminnelsen med en förseningsavgift. Men att jag dessutom tvingades betala 160 kronor extra i öppningsavgift tycker jag är upprörande. Vill bara att ni ska få veta.

Hälsningar,

Eva-Lena Hansson

C1. Varför skrev Eva-Lena Hansson detta brev?

- a) För att berätta att hennes mobil äntligen fungerade
- b) För att berätta att hon hade betalat för sent
- c) För att berätta att hennes mobil inte fungerade
- d) För att berätta att hon hade betalat för mycket

Som svar på frågan om frågan är relevant svarar 10 av 12 deltagare ”ja”. När deltagarna ska motivera sitt svar i enkäten menar Provkonstruktör 5 att det är ”relevant att kunna förstå syftet med ett meddelande.” Lärare 2 menar vidare i enkäten att:

Provtagaren ökar sin funktionella litteracitet genom att öva sig i denna typ av läsförståelse (det matchar läroplanen väl). Det är viktigt att kunna förstå klagomål samt framföra egna.
(Lärare 2)

10 av 12 deltagare valde svarsalternativet att frågan prövar förmågan att ”dra slutsatser utifrån information som uttrycks i olika delar av texten”. Att förstå syftet med en kommunikativ handling är en viktig del av den kommunikativa språkförmågan och något som ofta provas i test i kommunikativ språkförmåga. Det är också återkommande i de nationella proven i sf. Problemet i itemet i Figur 4 hade med svarsalternativen att göra: ”Relevant fråga men hittar inget rätt svar” (Provkonstruktör 6). I instruktionen till de automatiskt genererade frågorna ingick att endast ett svar skulle vara rätt. En fråga som ställdes i enkäten var därför ”Är endast ett svarsalternativ rätt eller kan flera vara rätt (vilka i så fall)?”

I relation till exemplet i Figur 4 ovan var uppfattningen olika. Det tänkta rätta svaret var alternativ d). Det finns dock problem med hur det rätta svaret är formulerat vilket några av deltagarna observerade i enkätsvaren:

Endast ett svarsalternativ är rätt men alternativ d också implicit, hon tycker iaf att hon har fått betala för mycket genom öppningsavgiften. Och det är detta faktum hon vill att den också ska veta. Så egentligen kan alternativ d också vara rätt förutom a, som är mer uppenbart (Lärare 6)

Inget, D är alldeles för otydlig och tolkningsbar. (Provkonstruktör 6)

Ett (Lärare 3)

Flera - alla? Felsvaren ligger väldigt nära rätt svar pga. lexikala överlapp - all info finns i texten. Därför svårt och lurigt om man inte vet att man får ange flera svar. (Provkonstruktör 5)

I resonemangen kring vad som behöver ändras för att flervalsfrågan ska bli mer välfungerande lyfter Lärare 2 följande i enkäten:

Angående svarsalternativet (d): "För att berätta att hon hade betalat för mycket", så kan vi inte veta om Eva-Lena verkligen har betalat för mycket. Det är en värdering – alltså inte objektiv information. Om vi istället formulerar svarsalternativ (d) såhär: "För att berätta att hon tycker att hon betalat för mycket", så anser jag däremot att det blir korrekt. Det är önskvärt att göra så, eftersom Eva-Lena uppenbarligen haft två anledningar till att skriva brevet: att informera om att mobilen fungerar igen och att hon anser sig ha betalat mer än vad hon borde (att företaget uppträtt girigt). (Lärare 2)

Skillnaden ligger alltså mellan uttrycket "hade betalat för mycket" och "tycker att hon betalat för mycket". Ändringarna som föreslås handlar således om ett behov av att komma närmare textens *kommunikativa syfte*.

Flera deltagare lyfter även att svarsalternativen är för långa och konstruktionen "att berätta" ställer till det eftersom det kan innefatta allt. Flera deltagare lyfter att de skulle vilja ha ett annat sorts svarsalternativ som är mer inriktat mot språkhandlingar. Som svar på frågan "Behöver något ändras i frågan så att den blir mer välfungerande? Vad i så fall?" lyfter några deltagare att "svarsalternativen bör ändras så att mer övergripande - typ "klaga", "tacka"" (Provkonstruktör 5, Provkonstruktör 2).

Pedagogiska möjligheter och begränsningar med att använda automatiskt genererade flervalsfrågor

Flervalsfrågor associerades generellt i intervjuerna i huvudsak med summativ bedömning. En del lärare lyfte också möjligheten att kunna arbeta formativt med flervalsfrågor genom att diskutera varför ett givet svar är rätt och hur man vet det. Det blir då ett sätt att arbeta med lässtrategier och ”då blir det en lärandesituation” (Lärare 2). Fördelarna som uppgavs riktades mot flervalsfrågor som en testteknik i prov. Det styrda formatet ”ger en möjlighet för alla att svara” (Provkonstruktör 5) och är inte beroende av deltagarnas skriftliga färdighet, vilket ansågs särskilt viktigt för kortutbildade sfi-elever. En annan fördel som uppgavs var möjligheten att ställa många frågor för att förbättra mätsäkerheten. Nackdelarna som lyftes var främst att det man begär elever att göra skiljer sig från läsning i icke-testsammanhang. ”Hur representativ är den här läsförståelsen i andra sammanhang? Det blir lätt väldigt konstruerat och lurigt” menar Provkonstruktör 1 och ser en ytterligare nackdel i att undervisningen riktas mot att lära eleverna att vara ”test wise”.

De främsta pedagogiska möjligheterna som lyftes i intervjuerna med just automatiskt genererade flervalsfrågor var möjlighet att få uppslag på frågor som man kanske inte själv hade tänkt på: ”när man sitter själv är man ju också begränsad med hur man brukar tänka och vad man är van vid /.../ [det] kan bidra till ytterligare ett sätt att se på det” (Lärare 5). Tekniken skulle alltså kunna fungera som en ”idéspruta” (Provkonstruktör 1). Provkonstruktör 2 beskriver hur ett system som genererar uppslag på frågor skulle kunna vara ett stöd i den kreativa processen som provkonstruktion innebär:

Det tar faktiskt tyvärr längre tid för mig nu att konstruera uppgifter än när jag var ny för jag har slut på idéer /.../ det känns motsägelsefullt att säga att jag blir sämre men jag upplever ju att jag blir det med det kreativa /.../ jag skulle vara hjälpt av det. (Provkonstruktör 2)

Att ta fram ett nationellt prov i sfi tar ca 12 månader (SOU 2022:17). I en provkonstruktionscykel är själva konstruerandet av uppgifter en mindre del. En stor del av tiden går åt till revidering, utprovning och möten med expertgrupper. Samtidigt som automatiskt genererade flervalsfrågor kan vara ett stöd i den kreativa processen ser Provkonstruktör 1 en farhåga i att generativ AI också kan ta över den del som många provkonstruktörer tycker är givande:

Många upplever den här konstruktionsfasen som kreativ och rolig, den är ganska spännande, att om man tar bort den och bara jobbar med revidering så är det inte så

jätteroligt /.../ man kanske hellre skulle vilja ha någon som gjorde alla revideringar.
(Provkonstruktör 1)

En annan aspekt som lyftes som en möjlighet var att automatiskt genererade flervälsfrågor skulle kunna vara tidssparande och att det blir ”lättare att kunna skapa sina egna prov” (Lärare 3) givet att tekniken blir välfungerande. Men den tidsbesparingen ansågs också vara avhängig hur mycket tid som behöver läggas på att justera frågorna så att de blir mer välfungerande. Lärare 1 menar att:

Man ser att man sparar tid men samtidigt då måste man kunna gå igenom dessa material och se är dom bra eller inte bra /.../ det räcker inte bara att du säger ja jag sparade tid fantastiskt men då måste jag också säkerställa att dom är bra frågor så att dom kan användas om jag vill använda dom som underlag. (Lärare 1)

De pedagogiska möjligheterna har också att göra med i vilket sammanhang tekniken är tänkt att användas. Så kallade high stakes prov, där mycket står på spel för individerna, ställer höga krav på frågor som fungerar ur ett provkonstruktionsperspektiv medan undervisningssammanhang öppnar upp för en mer flexibel användning. Provkonstruktör 1 såg möjligheten att använda automatiskt genererade flervälsfrågor i undervisningen som en fördel eftersom det blir ett sätt för lärare att enkelt skapa frågor till en text som elever kan få arbeta med, ett arbete som annars kan vara ganska tidskrävande. ”Det skapar ju en sorts aktivitet kring läsning som jag tror är intressant” men som också bygger på att ”läraren har en teoretisk idé om läsförståelse”. Att enkelt kunna generera olika typer av flervälsfrågor skulle också kunna öppna upp möjligheter till ökad individualisering:

Sen om vi jobbar mycket med digitalisering, det finns också ett bra sätt att träna med elever som kanske är väldigt snabba eller är väldigt duktiga och inte behöver så mycket stöd, dom kan göra det ensamma och träna med det. (Lärare 1)

Ett starkt önskemål från flera lärare var dock att det vid automatisk generering av flervälsfrågor ska gå att kunna önska svårighetsgrad och vilken typ av läsning som ska provas. Lärare 1 menar att det behövs ”lite mer frågor som också kräver att eleverna tänker så att det inte bara blir lokalisera utan också att dra slutsatser”.

När det gäller de begränsningar som lyftes i intervjuerna rörde det sig främst om att flervälsfrågorna ansågs vara alltför textnära och präglade av en hög grad av lexikal överlappning. Lärare 6 menade att några av flervälsfrågorna innehöll ”nonsensdistraktorer” och betonade att ”i förhållande till texten så behöver det

vara meningsfullt, det ska vara en relevans i innehållet i relation till texten, det ska inte vara plockat ur luften”. En av de största bristerna som lyftes fram av flera deltagare var att det inte alltid fanns endast ett rätt svar. Att kunna kontrollera det ansågs helt nödvändigt. ”Distraktionerna får inte vara så nära eller göra att det

blir så nära att man kan tolka in två svar som rätt”, menar Lärare 6 och betonar att det behövs ”en mänsklig hand i provkonstruktionen”. Ibland handlade det om enkla ändringar både i svarsalternativen och i frågekonstruktionen. Ett exempel som kom upp flera gånger i intervjuerna var frågan ”Vad är det som gör att många väljer att handla på internet?” som var kopplad till en faktatext. En sådan konstruktion öppnar för att många svar är rätt och det är svårt att avgöra var gränsen går för ”många”. Flera av deltagarna beskrev att om man ändrade till ”vilken är den vanligaste anledningen till att handla på internet” så fick frågan ett entydigt svar. ”Det är en liten ändring men den gör skillnad” (Lärare 5).

En annan konsekvens av att automatiskt generera flervalsfrågor är att de är framtagna i stunden och därmed inte är utprovade. Detta upplevdes som en begränsning då det kan uppstå en osäkerhet i vad de egentligen prövar:

Men det kan vara svårt då att bedöma just exakt och särskilt om dom inte är hundra procent testade förut och är dom automatiska så då har dom kanske inte varit testade förut så då har jag ingen möjlighet att veta vilka processer eller vilken kunskap eleven har. (Lärare 1)

Provkonstruktör 1 understryker i intervjun att även om resultatet vid ett prov som använder flervalsfrågor, automatiskt genererade eller ej, är lätt att få fram (t.ex. 28/38), är det en svårighet att tolka resultaten:

Att tolka resultaten /.../ då är det väldigt viktig när man kommunicerar det här vad det betyder. Vad visar den här AI:n, jag tänker att man kan ha det för underlag till diskussion men många kommer att vilja ha det här för att göra prov och så säger dom att man ska ha 70% rätt men vad betyder det? Det är jätteviktigt det måste finnas en tolkning av resultatet. (Provkonstruktör 1)

Dessa resonemang har en tydlig koppling till en central del av lärares bedömningskompetens, nämligen att kunna tolka och använda testresultat (t.ex. Fulcher, 2012; Pastore & Andrade, 2019). Det gäller för både mänskligt konstruerade uppgifter och automatiskt genererade, men det är möjligt att digitalisering av prov och en ökad förekomst av automaträttade uppgifter förstärker detta behov. Denna kunskap försvåras av det faktum att lärare och experter sällan är helt överens om vilken typ av läsning som bedöms (Tengberg,

2018). Provkonstruktör 1 har lång erfarenhet av att sitta i olika referensgrupper kopplade till standardiserade prov och menar att ”jag har suttit med i möten när man ska bestämma läsprocesser men vi var helt oense, det gick inte att enas” och att det har delvis att göra med att ”läsförståelse är en osynlig process i hjärnan”.

Om relationen mellan AI och mänskligt omdöme

Återkommande under intervjuerna lyfts behovet av att kunna granska automatiskt genererade frågor och att det behövs ”en mänsklig hand i provkonstruktionen” (lärare 6). Just förmågan att kunna granska och värdera lyfts i det som kallas ”AI literacy” (Ng m.fl., 2021). Det kan också ses som en viktig del i ett professionellt omdöme. Filosofen Jonna Bornemark menar att erfarenhet är en viktig del i det professionella omdömet (Bornemark, 2020). Provkonstruktör 2 preciserar erfarenhetens och omdömet roll i relation till generativ AI på följande vis:

Det mänskliga omdömet är att det här ser bra ut på pappret men jag vet vilka fem sex olika typ av missförstånd som brukar hända i klassrummet, dom kan jag förutse eftersom jag har erfarenhet av att vara i klassrummet /.../ AI har inte den erfarenheten så när den genererar en text med frågor så kan den inte förutse på vilka sätt man skulle kunna missförstå det här. (Provkonstruktör 2)

Provkonstruktör pekar på hur spåkmödelerna, som bygger på statistiska beräkningar, saknar erfarenhet utanför texten (se även Kuteeva & Andersson, 2024) vilket innebär en brist när det gäller att förutse frågornas funktionalitet i relation till en given målgrupp. Även Lärare 1 framhåller klassrumserfarenheten som central och att den bör samspela med teknologin:

Sen den kommer inte heller att kunna ersätta den erfarenhet som har lett till att vi har den här synen på undervisningen och på resultatet av vår undervisning och bedömning ehm så jag skulle säga det är fantastiskt att vi har möjlighet att kunna använda och ta nytta av teknologin för att göra nånting som är roligt och spännande och intressant men samtidigt så kan vi inte släppa allt det mänskliga för den erfarenhet som alla lärare och ni som jobbar med det /.../ och det krävs för att kunna se och ja kunna utvärdera det här. (Lärare 1)

Alla deltagare i studien är på olika sätt inne på att teknologin såsom den kom till uttryck vid studiens genomförande inte är tillräckligt nyanserad för att kunna användas utan en mänsklig blick. Att granska och vid behov justera automatiskt genererade flervalsfrågor sågs som centralt. Vilken typ av kunskap behöver då den som granskar ha? I intervjuerna lyfter deltagarna behov av kunskap om

provkonstruktion, kunskap om andraspråksutveckling och erfarenhet av målgruppen. Några av deltagarna lyfter att det förekom onödigt svåra och långa formuleringar i förhållande till målgruppens förväntade nivå. Som exempel nämns ”fått ut av” i frågan “Vad säger Louise att hon fått ut av mentorskapet?”, och svarsalternativet “leda” (som substantiv) och formuleringen “inbjuden”. Det fanns även exempel på onödigt långa formuleringar som i exemplet i Figur 4 där formuleringen ”för att berätta att” upprepas i varje svarsalternativ och i frågan ”Vad är det som gör att många handlar på internet?” i stället för bara ”Varför handlar...”. En särskilt viktig aspekt som framhålls i intervjuerna är behovet av omformuleringar för att komma närmare texternas kommunikativa syfte och för att endast ett svarsalternativ ska vara entydigt rätt. De flesta frågor bedömdes behöva en justering. Deltagarna är dock överens om att det i de flesta fall rör sig om små justeringar och betonar att det handlar om ”små små nyanser”(Lärare 4) och att de ”mänskliga glasögonen måste finnas med i dom här processerna” (Lärare 6).

På ytan kan automatiskt genererade flervalsfrågor se bra ut: de är grammatiskt korrekta, svarsalternativen är symmetriska och de anknyter till texten. Men när de granskas närmare framträder en annan bild. Som diskuterats ovan visar erfarna lärare och provkonstruktörer på hur automatiskt genererat språk behöver granskas och i många fall justeras för att uppfylla ett pedagogiskt syfte. Det finns därmed en fara i att använda automatiskt genererade flervalsfrågor oreflekterat. I en tid då alltmer text kommer att genereras automatiskt framstår behovet av expertkunskap och att kunna granska färdigformulerade förslag som viktigt. Som framkommit i den här studien består aktuell expertkunskap av kunskap och erfarenhet av målgruppen, provkonstruktion och andraspråksbedömning. Teknikutvecklingen för generativ AI går snabbt och det är svårt att förutsäga vilka karaktärsdrag automatiskt genererade flervalsfrågor kommer att ha i framtiden. Tekniken har dessutom sannolikt utvecklats sedan studiens genomförande. Oavsett den tekniska utvecklingen kommer lärare och provkonstruktörer att behöva kunna granska automatisk genererade flervalsfrågor för att kunna argumentera för vilka slutsatser som kan dras av elevers läsförståelse vid användning av automatiskt genererade flervalsfrågor för bedömning.

Avslutande diskussion

Generativa språkmodeller är inte längre en fråga bara för systemvetare och språkteknologer, utan berör allt fler i olika sektorer. Inte minst de verksamma i utbildningssektorn (Ng m.fl., 2021). Även om den specifika tekniken som använts i denna studie har utvecklats och kommer att utvecklas mer så är de övergripande frågeställningarna kring relationen mellan generativ AI och det mänskliga omdömet fortsatt relevanta. Detsamma gäller vilka konsekvenser som ny teknologi innebär för lärares och provkonstruktörers bedömningskompetens och för synen på kunskap. I en forskningsöversikt över studier om AI literacy i utbildningssammanhang är majoriteten av studierna ägnade åt hur AI literacy kan läras ut till elever (Ng m.fl., 2021). I den här studien diskuteras istället hur kunskap om generativ AI är en del av lärares bedömningskompetens.

Lärares bedömningskompetens påverkar kvaliteten på lärares bedömningar (Meijer m.fl., 2023). Flervalsfrågor lyfts ofta fram som exempel på ”objektiv” bedömning där det finns givna rätt och fel. Det är ett format som därmed är lätträttat och lämpar sig för digitala prov och automaträttade provdelar. Men hur resultat ska tolkas bygger också på en subjektiv uppfattning och vad en flervalsfråga egentligen prövar i relation till ett så pass komplext konstrukt som läsförståelse är inte helt givet (jmf. Tengberg, 2018). Bedömningskompetens kan ses som en del av ett professionellt omdöme för lärare som utvecklas över tid och i samspel med andra (Pastore & Andrade, 2019). Det professionella omdömet har således både en kollektiv och individuell dimension och utvecklas i ett samspel mellan kunskap och erfarenhet (Bornemark, 2020). Att kunna argumentera för och utveckla en samsyn bland kollegor kring vad läsförståelse innebär utifrån till synes ”enkla format” som flervalsfrågor är en viktig del av lärares bedömningskompetens i sammanhang där denna testteknik är vanligt förekommande. Tolkning och användning av provresultat lyfts i de flesta modeller för bedömningskompetens (t.ex. Fulcher, 2012; Pastore, 2023; Pastore & Andrade, 2019). Välgrundad tolkning och användning av bedömningsresultat är också en central del i bedömningars validitet (t.ex. Messick, 1989). Resultaten i föreliggande studie visar på hur ett format som är enkelt att rätta kan vara svårt att bedöma, när man med bedöma avser förmågan att kunna tolka och använda sig av resultaten. Det krävs en utvecklad bedömningskompetens för att kunna använda sig av informationen man får vid flervalsfrågor. En omvänd situation, där lärare inte kan tolka resultat innebär ett hot mot bedömningarnas validitet (Tengberg, 2018). Det är inte bara förmågan att tolka resultaten som är ett hot mot validiteten utan även själva frågornas kvalitet. Om de endast mäter så kallade

enklare läsprocesser fångar de inte det komplexa konstrukt som läsförståelse utgör i sin helhet. Frågorna riskerar då att utgöra en konstruktunderrepresentation (Messick, 1989).

I föreliggande studie har inte de automatiskt genererade flervälsfrågorna kunnat uppfylla ett pedagogiskt syfte fullt ut, bland annat genom att flervälsfrågorna var för textnära. En fråga för framtida forskning är hur språkmodeller kan utvecklas för bedömningssyften genom mer kvalitativ feedback från tilltänkta användare med pedagogisk kompetens. Även om språkmodellerna utvecklas (vilket de gjort sedan studiens genomförande) och kan följa instruktioner på ett bättre sätt än vad som var fallet i underlaget till denna studie, t.ex. genom att endast ett svarsalternativ är entydigt rätt och genom att undvika lexikal överlappning, kvarstår frågan om att kunna tolka och använda sig av resultat.

Automatiskt genererade frågor och automatiserad bedömning kan skapa en distans till den som blir bedömd. Men för de som skapar och på olika sätt använder sig av provresultat finns alltid ett ansvar när det gäller att kunna argumentera för hur provuppgifter ska tolkas och vilka slutsatser som kan dras vid bedömning. En utvecklad bedömningskompetens som inkluderar användning och kritisk granskning av automatiskt genererat språk blir särskilt viktigt i det testsamhälle vi lever i med en ökad förekomst av digitala test och standardiserade prov, både i utbildningssammanhang och i andra samhällssektorer. En utvecklad bedömningskompetens blir särskilt viktigt i sammanhang där konsekvenserna av bedömning riskerar att bli stora som exempelvis när språkbedömning föreslås ligga till grund för anställningsbeslut och för beslut i migrationsprocessen (Rydell, 2024; SOU 2024:78). Ytterligare frågor för framtida forskning är att följa användningen av språkmodeller i bedömningssammanhang, undersöka vilka konsekvenser lättillgänglig generativ AI får för synen på språklig kompetens och relevant kunskap samt etiska aspekter kopplade till generativ AI och bedömning.

Noter

2. I dagsläget gäller uppdraget om digitalisering av nationella prov inte de nationella proven i sfi. Sfi-proven var dock under en försöksperiod 2009-2011 digitaliserade.
3. Den här studien är en del av projektet *SWE-QUEST: Swedish Question Generation for Assessing Reading Comprehension* som finansierades av Digital Futures (2022-2024). I projektet deltog Johan Boye (KTH), Maria Rydell (SU) och Dmytro Kalpakchi (KTH).

Referenser

- Alderson, C. J. (2000). *Assessing reading*. Cambridge University Press.
- ALTE. (2005). *Material for the guidance of test item writers*. Association of Language Testers in Europe.
- Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2, 8-14.
- Bornemark, J. (2020). *Horisonten finns alltid kvar*. Om det bortglömda omdömet. Volante.
- Forsell, J. (2023). *AI inom bedömning väcker förhoppningar och farhågor*. Skolverket. <https://www.skolverket.se/skolutveckling/forskning-och-utvarderingar/artiklar-om-forskning/ai-inom-bedomning-vacker-forhoppningar-och-farhagor>
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9(2), 113–132.
- Institutionen för språkdidaktik. (2020). *Resultatrapport och uppföljning nationella prov i svenska för invandrare 2019 – kurs D*. Stockholms universitet.
- Kalpakchi, D. & Boye, J. (2023). Quasi: a synthetic question-answer dataset in Swedish using GPT-3 and zero-shot learning. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 477-491.
- Kalpakchi, D. & Boye, J. (2024). Generation and evaluation of multiple-choice reading comprehension questions for Swedish. *Northern European Journal of European Technology*, 10(1), 86-105.
- Kirsch, I.S. & Mosenthal, P.B. (1995). Interpreting the IEA reading literacy scales. I M. Binkley, K. Rust & M. Winglee (Red.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study* (s. 135–192). National Centre for Educational Statistics.

- Kuteeva, M. & Andersson, M. (2024). Diversity and standards in writing for publication in the Age of AI: Between a rock and a hard place. *Applied Linguistics* 45, 561-567.
- McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.
- Meijer, K., Baartman, L., Vermeulen, M. & de Bruijn, E. (2023). Teachers' conceptions of assessment literacy. *Teachers and Teaching*, 32(7-8), 695-709.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics. *Educational Researcher*, 18(2), 5-11.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Ng, D.T.K., Leung, J.K.L., Chu, S.K.W. & Qiao, M.S. (2021). Conceptualizing AI literacy: An explorative review. *Computers and Education: Artificial Intelligence*, 2, 100041.
<https://doi.org/10.1016/j.caei.2021.100041>
- OECD. (2019). *PISA 2018 Assessment and Analytical Framework*.
- Pastore, S. (2023). Teacher assessment literacy: a systematic review. *Frontiers in Education*, 8:1217167.
<https://doi.org/10.3389/feduc.2023.1217167>
- Pastore, S. & Andrade, H.L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, 84, 128-138.
- Popham, J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, 48, 4-11.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, 23(4), 441-474.
<https://doi.org/10.1191/0265532206lt337oa>
- Rydell, M. (2021). Bedömning på sfi. Artikel i Läslyftet *Stödstrukturer för undervisning på komvux*. Skolverket.
- Rydell, M. (2024). Språk som rättighet, skyldighet och projektionsyta för samtida debatter om migration. I D. Hedlund, D. Martinsson & K. Zillén (Red.), *Sammanhållning eller splittring: Olikgörande av barn och unga i samtidens Sverige* (s. 153-169). Jure Förlag.
- SKOLFS 2021:67. *Kursplan i kommunal vuxenutbildning i svenska för invandrare*.
- Skolverket (2020). *Redovisning av uppdrag om central rättning av nationella prov*. Dnr: 7.5.2.2019:1900.

- Skolverket (2024). Sätta betyg i grundskolan.
<https://www.skolverket.se/undervisning/grundskolan/betyg-i-grundskolan/satta-betyg-i-grundskolan#Text4>
- SOU 2022:17. *En modell för att mäta och belöna progression inom sf*. Statens offentliga utredningar. Utbildningsdepartementet.
- SOU 2024:78. *Ett språkrav för språkutveckling*. Betänkande från Utredningen om språkrav för personal i äldreomsorgen. Statens offentliga utredningar. Socialdepartementet.
- Tengberg, M. (2015). Att bedöma läsförmåga: Teoretiska utgångspunkter. I G-Skar & M. Tengberg (Red.), *Bedömning i svenskämnet. Årskurs 7-9*. (s. 87-108). Natur & Kultur.
- Tengberg, M. (2018). Validation of sub-constructs in reading comprehension tests using teachers' classification of cognitive targets, *Language Assessment Quarterly*, 15(2), 169-182.
<https://doi.org/10.1080/15434303.2018.1448820>
- Voss, E., Cushing, S.T., Ockey, G.J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly* 20(4-5), 520-532.
<https://doi.org/10.1080/15434303.2023.2288256>
- Weir, C. & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL Research Notes*, 31, 2-10.
- Wikström, C. (2013). *Konsten att göra bra prov: Vad lärare behöver veta om kunskapsmätning*. Natur & Kultur.
- Winerö, E. (2022). AI har dödat den skriftliga hemuppgiften. *Vi lärare*.
<https://www.vilarare.se/nyheter/vi-larare-debatt/ai-har-dodat-den-skrivna-hemuppgiften/>

Kontakt

Maria Rydell
maria.rydell@su.se