Eva Olsson
Göteborgs universitet
Sofia Nilsson
Göteborgs universitet
Anna Karin Lindqvist
Göteborgs universitet

DOI: http://dx.doi.org/10.5617/adno.6288

Test-taker feedback i utvecklingsprocessen av nationella prov i engelska

Sammanfattning
De nationella proven i engelska syftar till att stödja lärares betygssättning i svensk skola. Målsättningen i provutvecklingen är därför att konstruera prov med en så hög grad av validitet och reliabilitet som möjligt. Syftet med denna studie är, med utgångspunkt i Messick (1987, 1989), att undersöka och belysa på vilket sätt och i vilken utsträckning test-taker feedback kan bidra till provens validitet och reliabilitet. I en empirisk studie analyserades samvariation mellan elevers åsikter om läsförståelseuppgifter och det faktiska utfallet, det vill säga hur väl eleverna lyckades lösa uppgifterna. Data, som samlats in vid utprövning av nya uppgifter till det nationella provet i årskurs 9, bestod av feedback om nio läsförståelseuppgifter från cirka 400 elever per uppgift samt deras resultat på uppgiften. Analysen visar att elevers uppfattningar om hur bra uppgiften var, hur svår den var respektive hur väl de lyckades lösa den delvis samvarierar på ett statistiskt signifikant sätt med deras resultat när uppgiften poängsattes. Vidare visar resultaten att test-taker feedback kan tillföra värdefull information för att upptäcka om en uppgift tycks gynna någon grupp framför en annan. Informationen som test-taker feedback ger kan också bidra till stärkt validitet och reliabilitet om den exempelvis används för att sekvensera uppgifter utifrån upplevd svårighetsgrad eller för att sortera bort olämpliga uppgifter.

Nyckelord: Test-taker feedback, läsförståelse, nationella prov, engelska, validitet

The use of test-taker feedback in the development of national tests of english

Abstract
The purpose of the national tests of English is to provide support for teachers' grading of students in Swedish schools. Hence, the aim is to develop as valid and reliable tests as possible. Based on Messick (1987, 1989), the purpose of this study is to explore and illustrate in what ways and to what extent test-taker feedback may contribute to the validity and reliability of the tests. An empirical study was carried out, where the covariation between students' opinions about reading comprehension tasks and their actual results were analysed. Data consisted of test-taker feedback collected when trying out nine reading comprehension tasks for the national test in grade 9 among 400 students per task, and of students' results on the tasks. The analysis shows that the students' opinions about the overall quality and the difficulty of the tasks, as well as their outcome expectancy after completing the tasks, covaried in a statistically significant way with their performance, when the tasks were marked. Furthermore, the results indicate that test-taker feedback may provide useful information related to bias. The information from test-taker feedback may also contribute to the validity and reliability of a test, for instance when used for sequencing tasks according to experienced level of difficulty or for sorting out less suitable tasks.

Keywords: Test-taker feedback, reading comprehension, national tests, English, validity

Inledning

I Sverige genomförs obligatoriska nationella prov i engelska i grundskolans årskurs 6 och 9 samt i den högsta avslutande kursen för respektive gymnasieprogram, det vill säga kursen Engelska 5 eller Engelska 6 beroende på vilket program det gäller.[1] De nationella proven har, enligt Skolförordningen (SFS 2017:1107), till syfte att stödja betygssättningen. Proven är inte examensprov utan en del av det underlag lärare använder när slutbetyg sätts. Dessutom finns icke-obligatoriska bedömningsstöd i engelska för andra årskurser/kurser samt i moderna språk (spanska, tyska och franska) för grundskolan och gymnasieskolan. De nationella proven och bedömningsstöden i engelska och moderna språk konstrueras på uppdrag av Skolverket av projektet *Nationella prov i främ-*

[1]Se Skolverkets webbplats https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/provdatum-i-gymnasieskolan

mande språk (NAFS) vid Institutionen för pedagogik och specialpedagogik vid Göteborgs universitet.[2]

En viktig målsättning i provverksamheten är att utveckla prov med en så hög grad av validitet och reliabilitet som möjligt. Grundläggande utgångspunkter för all typ av bedömning i skolan bör, enligt Erickson (2016), vara tydlighet, giltighet, tillförlitlighet och respekt – begrepp som förknippas med validitet och reliabilitet (jfr Bachman & Palmer, 1996; Bachman, 2000; Kane, 2002; Moss, 2007). I linje med Messick (1987, 1989) poängterar Little och Erickson (2015) att transparens, etik och respekt i förhållande till de elever vars kunskaper bedöms är centralt för validiteten och reliabiliteten i bedömningen. De framhåller att elever bör förstå vad som ska bedömas samt att uppgifter som ges på ett tydligt sätt ska svara mot detta konstrukt och vara begripliga för de elever som genomför dem (jfr Kunnan, 2000). Erickson (2010) understryker att elever och deras lärare utgör viktiga expertgrupper i konstruktions- och valideringsprocessen av prov och uppgifter eftersom de kan ge information om hur uppgiften fungerar ur deras perspektiv (se även Erickson & Åberg, 2012). Forskning om hur elever uppfattar prov och uppgifter kan således tillföra kunskap som är högst relevant vid provkonstruktion, vilket exempelvis poängterats av Cumming (2004) och Ryan (2014), eftersom insikter om faktorer som kan påverka utfallet är betydelsefulla av reliabilitets- och validitetsskäl.

Syftet med denna studie är att undersöka och belysa på vilket sätt och i vilken utsträckning *test-taker feedback* – synpunkter från elever som prövar ut uppgifter – kan tänkas bidra till de nationella provens reliabilitet och validitet. I en begränsad empirisk studie undersöks i vilken utsträckning elevers åsikter om ett antal läsförståelseuppgifter samvarierar med det faktiska utfallet, det vill säga hur väl eleverna lyckades lösa uppgifterna. Vidare diskuteras hur den information som test-taker feedback ger kan användas i vidareutvecklingen av uppgifter. I undersökningen används feedback insamlad i samband med utprövningar av tre typer av läsförståelseuppgifter till det nationella provet i engelska för årskurs 9 för att exemplifiera och problematisera användningen av test-taker feedback i konstruktionsprocessen. Följande fråga är i fokus: På vilket sätt och i vilken utsträckning kan test-taker feedback bidra till att stärka de nationella provens reliabilitet och validitet?

Tidigare forskning

Vad som ska bedömas i skolan, vem som ska bedöma samt när och hur bedömningen ska ske är centrala frågor för enskilda individer, såsom elever och lärare. Det är emellertid också viktiga frågor ur ett samhällsperspektiv, bland annat

[2] Se NAFS webbplats https://nafs.gu.se

eftersom bedömning kan ha stor inverkan på hur utbildning utformas och vem den görs tillgänglig för (Fulcher, 2009; Korp, 2011). Synen på bedömning i Sverige har, enligt Erickson (2016, s. 4), utvecklats ”från ett i huvudsak tekniskt och mätrelaterat till ett mera pedagogiskt perspektiv” under de senaste decennierna. Erickson menar att intresset för bedömningsfrågor i allmänhet tycks ha ökat under denna period, vilket bland annat kan bero på den ökade fokuseringen på mål och kunskapskrav i de senaste läroplanerna (se Skolverket, 1994, 2011).

När bedömningens pedagogiska syfte poängteras ses bedömningen som en del av undervisningen och den används ofta i *formativt syfte*, det vill säga resultatet av bedömningen blir en utgångspunkt för fortsatt undervisning och lärande (Black & Wiliam, 1998, 2018; Hattie & Timperley, 2007; Hirsh & Lindberg, 2015). Självbedömning, kamratbedömning och återkoppling från läraren inför fortsatt bearbetning av en uppgift eller inför nästa uppgift är typiska inslag i formativ undervisning. När läraren exempelvis inför betygssättning ska bedöma elevens kunskaper på ett summerande sätt i förhållande till kunskapskraven sker bedömningen i *summativt syfte*. Ibland ses bedömning i formativt respektive summativt syfte som varandras motsatser, men man kan också hävda att all formativ bedömning börjar i någon typ av summering av vad eleven redan kan (Taras, 2005; jfr Klapp, 2015).

Nationella prov i Sverige har i huvudsak ett summativt syfte, det vill säga de syftar till att ge en bild av elevers kunskapsnivå vid en viss tidpunkt, men givetvis kan proven också användas i formativt syfte (Erickson, 2010). Både lärare och elever får genom provet en bild av elevens kunskaper i det aktuella ämnet, vilket kan vara ett stöd för fortsatt planering av undervisning och lärande. Det primära syftet är dock att pröva elevers kunskaper mot de kunskapskrav som finns i den aktuella årskursen och att stödja en likvärdig och rättvis bedömning genom standard-setting (Gustafsson, Cliffordson & Erickson, 2014). I ett nationellt prov prövas således elevernas kunskaper mot en gemensam måttstock, nämligen kunskapskraven för årskursen, medan undervisningen i större utsträckning kan anpassas utifrån den nivå där eleven är; ett förhållande som kan ses som ett dilemma (jfr t.ex. Nilholm, 2005). En elev som gör sitt bästa utifrån sina förutsättningar men inte lyckas nå godkänd nivå kan tappa motivationen för fortsatt lärande (Klapp, Cliffordson & Gustafsson, 2014; Klapp, 2015).

I all hantering av prov och bedömning, vare sig det gäller övergripande policynivå eller i praktiken i klassrummet, och inte minst i arbetet med att utveckla storskaliga prov, finns etiska aspekter att överväga, vilket inte minst Messick (1989) framhåller som centralt i valideringsprocessen. I Kunnans (2004) *Test Fairness Framework*, lyfts bland annat opartiskhet och konsekvenser av prov som faktorer som bör undersökas när man bedömer hur rättvist och rättvisande ett prov är. Kunnan vänder sig mot traditionella, snävare sätt att se på validitet och reliabilitet och menar liksom Zieky (2006) att en *fairness review* bör göras för att stärka ett provs validitet.

Shohamy (2001; 2017) uttrycker att både de som konstruerar och de som använder prov har ett ansvar för att skydda de elever som åläggs att genomföra prov; elever ska inte fara illa av att genomföra dem. Hon vidhåller att prov behövs och är användbara men också att syftet och användningen av prov ständigt måste övervägas så att provet blir en del av lärprocessen och inte ett hinder (Shohamy, 2017). I linje med Messick (1989), Shohamy (2001) och Bachman (2000) poängterar även Cheng och DeLuca (2011) vikten av att väga in konsekvenserna av storskaliga prov i validitetsprövningen, eftersom de kan inverka på individens framtid, exempelvis vad gäller val av utbildning och yrke, därför att individens självbild och självkänsla kan komma att påverkas. Cheng och DeLuca menar att deltagarperspektivet är centralt i valideringen och att det är nödvändigt att det tas tillvara (jfr Fox & Cheng, 2015; Spolsky, 2017).

En elevs inställning till ett prov påverkar givetvis resultatet liksom elevens strategier när provet genomförs (se t.ex. Purpura, 1999; Cohen, 2007). En elev som känner osäkerhet eller oro riskerar att prestera sämre än vanligt, vilket i sig tydliggör nödvändigheten av att i valideringen inbegripa aspekter som har med deltagarnas uppfattningar att göra (Messick, 1989; Elder, Iwashita & McNamara, 2002). Kunnan (1994) lyfter fram att faktorer som kulturell och språklig bakgrund, kön, tidigare skolgång, motivation och personlighet kan påverka hur deltagare lyckas på prov och att provutvecklare måste visa en medvetenhet om detta (se även Huhta, Kalaja & Pitkänen-Huhta, 2006).

Om ett prov ska ge rättvisande resultat bör eleverna ges så stor möjlighet som möjligt att visa vad de kan snarare än vad de inte kan. Att därför låta elever medverka i valideringsprocessen kan, enligt Little och Erickson (2015), sägas utgöra en demokratisk aspekt av provutvecklingen, en fråga om respekt, när de som är mottagare av proven också får vara med och påverka dem (jfr Alderson, Clapham, & Wall, 1995; Shohamy, 2001; Bachman & Palmer, 2010; Ryan, 2014). Att det finns en acceptans för de uppgifter som används i prov bland elever och lärare, att elever upplever att provet ger dem möjlighet att visa vad de kan och att innehållet uppfattas som relevant kan således ses som viktigt ur validitetssynpunkt (Erickson, 2010; Erickson & Åberg, 2012).

Ibland kallas den aspekt av validitet som berör deltagarnas grad av acceptans av prov för *face validity* (se t.ex. Cronbach, 1984; Nevo, 1985) men som påpekats av exempelvis Secolsky (1987) kan givetvis ett tests eller ett items validitet inte enbart prövas genom att analysera test-taker feedback. Andra typer av kvantitativa och statistiska analyser kan snarare kompletteras genom analys av test-taker feedback, för att sammantaget ge en bild av konstruktets validitet (jfr Messick, 1989).

Att test-taker feedback kan bidra till stärkt validitet fann exempelvis Brown (1993) i en studie som gällde ett japanskt språktest. Bland annat kunde informationen om provet utvecklas så att deltagarna tydligare förstod syftet med provet och hur det genomfördes eftersom feedback visade att en grupp studenter tycktes missförstå och därmed missgynnas. Enligt Brown är systematisk analys av

test-taker feedback under konstruktionsfasen viktig eftersom provets kvalitet därigenom kan förstärkas (se även Xie, 2011, Stricker, 2012; Huang & Hung, 2017).

Även Ryan (2014) framhåller att test-taker feedback ger kvalitativ information om ett prov eller en uppgift som kompletterar den kvantitativa analysen. Genom test-taker feedback kan man få inblick i deltagarnas positiva och negativa erfarenheter i samband med genomförandet. Enligt Ryan är denna information värdefull i konstruktionsarbetet eftersom vissa, annars svårfångade, aspekter av ett prov eller en uppgift kan framträda, såsom faktorer som underlättar respektive försvårar möjligheten för deltagare att visa vad de kan (se även Bachman & Palmer, 1996). Ryan påpekar också att deltagare även kan komma med konkreta och användbara förslag på hur ett prov eller en uppgift skulle kunna förbättras vilket innebär att feedback kan vara en viktig del i det ständiga arbetet att förbättra prov och provsystem.

I föreliggande studie är test-taker feedback om läsförståelseuppgifter i fokus. Rupp, Ferne och Choi (2006) undersökte hur olika läsförståelseuppgifter uppfattades av deltagare som genomförde prov. Resultatet visade bland annat att flervalsfrågor uppfattades pröva problemlösning snarare än läsförståelse. Rupp et al. menar att designen av enskilda frågor och valet av texter är det som styr hur provet uppfattas och vad det verkligen testar. I en studie av Sasaki (2000) visade sig även deltagarnas bakgrundskunskap om de ämnen som lucktexter i läsförståelseprov innehållsmässigt behandlade ha betydelse för hur deltagarna tog sig an uppgifterna och hur de klarade dem.

Det finns alltså ett stort antal faktorer som kan och bör vägas in under en valideringsprocess. Messick (1989, s.13) uttryckte följande:

> "Validity is 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".

Med andra ord beror graden av validitet i ett prov eller i en uppgift på hur väl underbyggda de slutsatser man drar utifrån resultatet är. I denna studie belyses hur test-taker feedback kan bidra i valideringsprocessen.

Bakgrund

För att kontextualisera den föreliggande studien ges en kortfattad beskrivning av det nationella provet i engelska i årskurs 9. Något mer detaljerat beskrivs den del av provet som avser pröva läsförståelse eftersom studiens empiri är hämtad från utprövning av sådana uppgifter.

Det nationella provet i engelska för årskurs 9

Provet i årskurs 9 är det sista nationella provet i engelska i den obligatoriska grundskolan vilket innebär att ett stort antal elever genomför det. Exempelvis genomförde cirka 93 000 elever provet läsåret 2016/2017. I årskurs 9 är de flesta elever 15–16 år gamla och de har vanligtvis studerat engelska i skolan sedan lågstadiet[3].

Vad de nationella proven i engelska ska pröva och hur processen att ta fram reliabla och valida prov utformas, fastställs formellt i ett provramverk av NAFS och Skolverket i samråd, där Skolverket fastställer yttre ramar på systemnivå[4] och NAFS utarbetar ramverk för de enskilda provens innehåll och konstrukt. En noggrann analys av kursplanen i engelska och de kunskapskrav som anges för årskurs 9 i Läroplan för grundskolan samt för förskoleklassen och fritidshemmet (Skolverket, 2011) utgör grunden för provutvecklingen. Målsättningen är att skapa prov vars resultat, så långt möjligt, återspeglar elevens faktiska kunskapsnivå i språket, men all kunskap och alla förmågor en elev besitter i ett språk kan givetvis inte rymmas i ett prov.

Det nationella provet i engelska för årskurs 9 består, i likhet med provet för årskurs 6 och proven i gymnasieskolan, av tre delprov, varav ett är uppdelat i två delar. Delprov A avser pröva muntlig produktion och interaktion, delprov B receptiv kompetens uppdelat på läs- och hörförståelse, och delprov C skriftlig produktion och interaktion. Delprovens uppgifter är tänkta att svara mot de målformuleringar som finns i kursplanen för engelska vad gäller dessa kompetenser (Skolverket, 2011). Varje delprov har alltså ett huvudsakligt fokus på en förmåga men delvis prövas förmågorna integrerat, inte minst därför att autentiskt språkbruk ofta innebär att de integreras. Svarsformaten varierar inom ett prov. I delprov A och C ges svar i form av ren produktion och interaktion, det vill säga eleverna deltar i samtal respektive skriver en text, ofta till en tänkt mottagare. I delprov B, där uppgifterna utgår från inspelat material respektive skrivna texter av olika slag, ingår frågor där eleven själv formulerar svar i skrift men även uppgifter av flervalstyp, där eleven väljer bland flera svarsalternativ eller exempelvis matchar en text med en rubrik. Dessa båda svarsformat förekommer i ungefär lika stor omfattning i Delprov B, vilket ger elever möjlighet att visa sin receptiva förmåga även på annat sätt än genom att själva formulera skriftliga svar.

Innehållet i uppgifterna väljs utifrån kursplanens centrala innehåll, vilket för årskurs 9 innebär att de exempelvis kan anknyta till vardagliga situationer, intressen, händelser, känslor, åsikter, sociala relationer, etiska frågor och levnads-

[3] Skolans styrdokument anger inte exakt under vilken av årskurserna 1–3 engelskundervisningen ska starta utan enbart hur många timmar som ska läggas ut under lågstadiet.

[4] Se Skolverkets systemramverk för nationella prov: https://www.skolverket.se/publikationer?id=3890

villkor. Exempel på uppgiftstyper som kan ingå i de olika delproven finns publicerade på projektet NAFS webbplats.[5]

Läsförståelse i det nationella provet

I delprov B förekommer olika typer av texter och uppgifter för att täcka in skilda aspekter av läsförståelse (se t.ex. Grabe, 2009; Grabe & Stoller, 2011). I kursplanens precisering av innehåll i engelska för årskurs 7–9 nämns i samband med läsning, att undervisningen ska syfta till att utveckla elevernas förmåga att förstå och tolka innehållet i olika slags texter såsom fiktion, beskrivningar, information, nyheter och reportage (Skolverket, 2011). Eleverna ska också, enligt kursplanen, utveckla språkliga strategier för att förstå och uppfatta såväl detaljer som sammanhang i texter. Det kan exempelvis handla om lässtrategier för att hitta viss information eller för att förstå en längre sammanhängande berättelse (jfr Grabe, 2009; Börjesson, 2012). I samband med såväl receptiv som produktiv förmåga nämns i kursplanen att undervisningen också ska syfta till att utveckla kännedom om grammatiska strukturer, satsbyggnad, ord med olika stilvärden samt fasta uttryck i språket. Detta innebär att det delprov som syftar till att pröva läsförståelse kan innehålla uppgifter där exempelvis kännedom om satsbyggnad och fasta uttryck prövas.

Uppgiftstyper som prövar läsförståelse

Delprovet som avser pröva läsförståelse består vanligtvis av tre till fyra olika uppgifter. Med *uppgift* avses här en text med frågor som anknyter till texten eller en text med luckor där ord ska markeras eller skrivas in. Varje fråga eller lucka kallas här *item*. En uppgift kan innehålla 6–25 items beroende på uppgiftstyp.

En av uppgifterna i delprovet består av en längre, berättande text om cirka 1200–1400 ord, en så kallad *lång läsförståelse (LL)*, där förmåga att läsa, förstå och i viss mån tolka längre sammanhängande text prövas, vilket exempelvis innebär att läsaren behöver kunna dra slutsatser även om sådant som inte explicit sägs i texten. Uppgiftstypen LL innehåller både öppna frågor, där eleven själv producerar ett svar, och frågor med flervalsalternativ, där eleven markerar ett av fyra till fem alternativ. Totalt ingår vanligtvis 20–25 items i en uppgift av LL-typ.

Någon av de kortare uppgifterna i delprovet utgörs ofta av en så kallad *luck-text*, en text där ett antal strategiskt valda ord tagits bort och ersatts med luckor (se Bachman, 1985; Velling Pedersen, 2009). I uppgiftstypen *multiple cloze (MC)* väljer eleverna ett ord som passar in i luckan bland fyra till fem alternativ:[6]

[5] www.nafs.gu.se
[6] Exempel från www.nafs.gu.se

This is a text about the city of Toronto in Canada, a _____ where people from all over the world have come to settle.	A museum B building C school D place E continent

En annan typ av lucktext är *open cloze (OC)*, där inga alternativ finns givna utan eleven måste själv komma på och skriva in ett ord som passar i sammanhanget:

This is a text ____*about*____ native languages.

Open cloze har alltså ett starkare inslag av produktion än multiple cloze eftersom eleven själv måste fylla i de ord som saknas.

Lucktexter förekommer dels med ett längre, sammanhängande textunderlag om 300–500 ord med 12–16 luckor, dels i form av fristående, korta dialoger där varje dialog innehåller enstaka luckor. Uppgifter i form av lucktexter avser att ge elever möjlighet att visa sin förmåga att förstå sammanhang, att språkligt binda ihop innehåll och att exempelvis visa kännedom om fasta uttryck och grammatiska strukturer. Som exemplen visar är de utelämnade orden oftast relativt vanligt förekommande ord; avsikten är inte att specifikt pröva ordförrådets djup eller bredd, utan förmågan att binda ihop innehåll. Det strategiska urvalet av luckor (se Bachman, 1985) innebär att eleverna ibland kan fylla i en lucka enbart med ledning av den sats där luckan finns. Andra luckor kräver förmåga att binda ihop text utifrån en något vidare kontext, såsom med hjälp av en annan sats i samma mening eller en närliggande mening. Vissa luckor, de som Bachman hävdar är svårast, innebär att eleven måste förstå vilket ord som ska fyllas i luckan genom att tänka utanför själva texten (se även Chapelle & Abraham, 1990; McCray & Brunfaut, 2018).

Ytterligare en uppgiftstyp som förekommer i delprovet är en så kallad *Info seek* där eleverna ska översikts- och lokaliseringsläsa flera kortare texter, ofta kring ett gemensamt tema, för att besvara dels öppna frågor, dels frågor av flervalstyp, exempelvis genom att kombinera ett antal påståenden med rätt textavsnitt.

I den empiriska undersökning som genomförs i denna studie analyseras testtaker feedback om tre av de ovan nämnda uppgiftstyperna, nämligen lång läsförståelse, multiple cloze och open cloze. Dessa uppgiftstyper används regelbundet i provet för årskurs 9 liksom i andra nationella prov i engelska på olika nivåer. De representerar uppgiftstyper som är tydligt olika till sin karaktär (lång

läsförståelse och lucktexter) men också uppgiftstyper som på ytan kan synas likartade men som kan uppfattas olika av elever (multiple cloze och open cloze).

Utprövningsprocess

I utvecklingsarbetet av nya provuppgifter, vilket sträcker sig över en tidsrymd om cirka två år, ingår en omfattande utprövningsprocess, där ett stort antal elever och deras lärare medverkar. Utprövningen utgör en viktig del av valideringen av uppgifter och prov; sedan 1990-talet har test-taker feedback samlats in som en del av valideringsprocessen i samband med utvecklingen av språkprov vid Göteborgs universitet (Erickson, 1999).

Till att börja med testas en uppgift i ett fåtal klasser för att få en uppfattning om hur väl uppgiften fungerar och på vilket sätt den behöver utvecklas. Både elever och undervisande lärare i de klasser som deltar i dessa mindre utprövningar ombeds lämna synpunkter om uppgifterna. Om exempelvis analysen av synpunkterna eller det faktiska utfallet visar att någon uppgift eller specifikt item verkar vara för lätt eller för svårt, eller kanske otydligt eller missvisande, bearbetas uppgiften. Även synpunkter om texternas innehåll tas tillvara. Om en stor andel av de elever och lärare som deltar i mindre utprövningar ogillar innehållet i någon text som ingår, lyfts den ur processen för att aldrig återkomma eller, om det bedöms möjligt och lämpligt, för att bearbetas och åter prövas ut.

Efter dessa mindre utprövningar och justeringar – en uppgift prövas ofta ut i mindre skala i flera omgångar – skickas uppgifter som bedöms ha potential att fungera väl ut till stor utprövning. Cirka 400 elever vid slumpvis utvalda skolor, oftast två klasser per deltagande skola, genomför samma uppgifter. I ett utprövningshäfte ingår oftast en till tre uppgifter, beroende på hur tidskrävande de är.

Vid dessa storskaliga utprövningar ombeds elever och lärare lämna omdömen om uppgifterna, och den data, test-taker feedback, som därigenom genereras ingår i den efterföljande analysen. Elevens senaste betyg i engelska anges av läraren för att användas i analysen av provets reliabilitet och validitet för bedömning av olika kunskapsnivåer.

Nedan visas hur det elevformulär för test-taker feedback som ingår i utprövningsmaterialet kan vara utformat:

We need YOUR help to make really good tests of English.

*Please react to the following statements about **XXX***
and then write your comments.

		Yes, absolutely				No, absolutely not
1	*XXX* was a good test	____	____	____	____	____
2	It was difficult	____	____	____	____	____
3	The text was interesting to read	____	____	____	____	____
4	There were many words that I didn't understand	____	____	____	____	____
5	I think I did well on this part of the test	____	____	____	____	____

Comments about *XXX* *(you can write in English or in Swedish)*

_____

_____

Figur 1. Elevenkät

Som elevformuläret visar ombeds eleverna uttrycka synpunkter om provuppgifterna på en femgradig Likertskala. Frågorna har varierat något genom åren och kan skilja sig åt beroende på uppgiftstyp. Eleverna ombeds dock alltid ange hur bra de tycker en uppgift är, hur svår de anser att den är och hur väl de upplever att de klarat den aktuella uppgiften. Därför används dessa tre frågor när feedback analyseras i föreliggande studie. Eleverna ges också möjlighet att i enkäten skriva kommentarer fritt.

I utprövningar ingår alltid, förutom nya uppgifter, även en så kallad ankaruppgift. Det är en uppgift som använts i utprövningar i årskurs 9 under lång tid – i dagsläget finns data från cirka 20 000 elever. Ankaruppgiftens funktion är att länka mellan år och utprövningsgrupper, och den används vid analys och utvärdering av utprövningsmaterialet. Lösningsfrekvensen för ankaruppgiften, det vill säga den andel av uppgiftens delfrågor, items, som en utprövningsgrupp i genomsnitt klarat, kan jämföras med resultat för andra utprövningsgrupper som genomfört ankaruppgiften vid andra tillfällen. Denna indikation om gruppernas nivå är betydelsefull information i analysarbetet av nyproducerade uppgifters svårighetsgrad.

Efter den stora utprövningen analyseras resultaten av olika uppgifter på detaljnivå och utfallet av varje item undersöks för att säkerställa att uppgifter av olika svårighetsgrad finns med och för att så långt möjligt säkerställa uppgifternas reliabilitet och validitet. En kombination av kvalitativa och kvantitativa metoder används i dessa analyser av utprövningsdata. De kvantitativa analyserna baseras i huvudsak på den klassiska mätläran. Exempelvis analyseras uppgifters medelvärde och spridning, liksom diskriminationsförmåga (R-bis, R-p-bis, CITC), mätfel och reliabilitet (K-R 20/21, Cronbachs alpha). I de kvalitativa analyserna av utprövningsdata ingår noggrann granskning och bedömning av elevsvar, till exempel vilka ord eleverna skrivit i luckor eller hur de har besvarat frågor om texten, vilket tillsammans med analysen av test-taker feedback och synpunkter från lärare kan föranleda förändringar i uppgiften.

Ett antal nya uppgifter prövas ut varje år och de som utprövningen visar fungerar väl samlas i en uppgiftsbank för att användas i olika kombinationer i kommande prov. En väl fungerande uppgift kan komma att återanvändas i prov en till tre gånger under den period den är sekretessbelagd för att länka mellan prov.[7]

Metod

I detta avsnitt beskrivs den data och de analysmetoder som använts i den föreliggande studien för att undersöka och belysa på vilket sätt och i vilken utsträckning test-taker feedback kan bidra till att stärka de nationella provens reliabilitet och validitet.

I en empirisk undersökning analyserades resultat och feedback från utprövningar av läsförståelseuppgifter till provet i årskurs 9. Utprövningsdata som gällde tre olika uppgiftstyper för att pröva receptiv förmåga avseende läsförståelse har använts: lång läsförståelse (LL), multiple cloze (MC) och open cloze (OC). I studien ingår resultat och test-taker feedback från utprövning av nio uppgifter, tre av varje uppgiftstyp. Urvalet består av utprövningsdata från de tre senaste uppgifterna av respektive uppgiftstyp som efter utprövningen använts i skarpa prov (t.o.m. provet 2018). Eftersom uppgifterna kan återanvändas kan en uppgift som nyligen förekom i ett prov ha prövats ut för ett antal år sedan. Var och en av de nio uppgifterna som ingår i studien har prövats ut av en elevgrupp om cirka 400 elever i årskurs 9 vid ett utprövningstillfälle någon gång mellan 2003 och 2013. I Tabell 1 visas hur många elever som prövat ut varje uppgift.

[7] Skolverket fastslår sekretessperiodens längd, vanligtvis är den 6 år. Ibland kan copyright-restriktioner innebära att en uppgift bara kan användas under en begränsad tid eller i ett enda prov.

Tabell 1. Antal elever per utprövad uppgift

Uppgift	LL1	LL2	LL3	MC1	MC2	MC3	OC1	OC2	OC3
Antal elever	498	488	465	409	424	358	389	400	431

Eftersom uppgifterna kan ha bearbetats något efter den stora utprövningen och eftersom test-taker feedback inte brukar samlas in från elever som genomför ett skarpt nationellt prov, görs i denna studie inga jämförelser med utfall vid skarpt prov.

Eleverna besvarade i samband med utprövningen en enkät (se Figur 1), där de på en femgradig Likertskala tog ställning till olika påståenden om uppgiften. I studien har svaren som gällde tre påståenden om uppgiften använts, eftersom dessa tre påståenden fanns med vid alla utprövningstillfällen:

”X was a good test” (X står för titeln på den aktuella uppgiften)
“It was difficult”
“I think I did well on this part of the test”

När data som är baserad på elevernas angivelser på den femgradiga skalan analyseras står värdet 5 för det mest positiva svaret, det vill säga eleven anser att uppgiften är bra, att den inte är svår respektive att eleven anser sig ha klarat uppgiften bra. Värdet 1 står för det motsatta.

I analysen har medelvärden för elevernas feedback, baserade på den femgradiga skalan, sammanställts för var och en av de ingående uppgifterna. Vidare har statistiska korrelationsanalyser (Pearson) använts för att jämföra elevernas uppfattning om uppgiften med det faktiska utfallet när uppgiften poängsatts. Även samvariationen mellan elevernas värdering av uppgiften – om den var bra eller dålig – och uppfattningen om hur väl de lyckats har undersökts. Pojkars och flickors feedback har också jämförts.

Begränsningar

Eftersom endast feedback från nio uppgifter ingår i studien, tre av vardera uppgiftstyp, går det inte att generalisera utifrån resultat där exempelvis uppgiftstyper jämförs, men resultaten kan ändå vara av intresse eftersom sammantaget ett stort antal elever ingår i studien. Resultaten används främst för att belysa vilken typ av information test-taker feedback kan ge och hur den kan användas för att stärka provens reliabilitet och validitet.

När test-taker feedback tolkas bör en medvetenhet finnas om de psykologiska faktorer som kan spela in när en elev väljer att markera ett visst alternativ på en Likertskala (se t.ex. Launeanu & Hubley, 2017). Elevens självbild och självförtroende har givetvis stor inverkan när eleven ska bedöma hur bra hen

klarade en viss uppgift och det finns också ett tolkningsutrymme för vad skalans steg innebär.

Vid tolkning av resultaten bör det tas i beaktande att varje uppgift är utprövad av en grupp om cirka 400 elever. Alla nio uppgifter har alltså inte prövats ut av samma elever. Vidare påminns om att det ingår olika uppgiftstyper i materialet vilket kan ha betydelse vid tolkning av samvariation mellan uppfattningar om uppgifter och lösningsfrekvens.

Resultat och analys

Inledningsvis redovisas en kvantitativ sammanställning av test-taker feedback för var och en av de nio läsförståelseuppgifterna. Därefter följer jämförande analyser av elevers uppfattningar om uppgifterna och det faktiska utfallet.

Medelvärden per uppgift

Tabell 2 visar medelvärden för test-taker feedback där eleverna i varje utprövningsgrupp på en femgradig Likertskala tagit ställning till hur bra de ansåg att den utprövade uppgiften var, hur svår den var samt i vilken grad de upplevde att de klarat uppgiften. Ett högt medelvärde betyder att gruppen i genomsnitt är mer positivt inställd än vid ett lägre medelvärde. Tabellen visar test-taker feedback för tre långa läsförståelseuppgifter (LL 1–3), tre multiple cloze-uppgifter (MC 1–3) och tre open cloze-uppgifter (OC 1–3). Den genomsnittliga lösningsfrekvensen och ett reliabilitetsvärde (Cronbachs alpha) redovisas för var och en av de uppgifter som prövats ut liksom för ankaruppgiften[8], en uppgift som var densamma i alla utprövningar.

[8] Ankaruppgiften bestod av en lucktext i form av korta dialoger om ett par repliker med en lucka per dialog. Den innehöll 12 items.

Tabell 2. Test-taker feedback, lösningsfrekvens och Cronbachs α för utprövade uppgifter samt ankaruppgift

	Utprövad uppgift (antal items)								
	LL1 (21) m	LL2 (23) m	LL3 (21) m	MC1 (15) m	MC2 (13) m	MC3 (13) m	OC1 (11) m	OC2 (12) m	OC3 (19) m
"It was a good test"	3.8	3.3	3.5	3.7	3.8	3.8	3.3	3.6	3.7
"It was difficult"	3.1	2.4	2.6	3.1	3.1	3.2	2.8	2.7	2.9
"I think I did well"	3.4	3.2	3.1	3.3	3.4	3.5	3.1	3.1	3.3
Lösningsfrekvens utprövad uppgift	.66	.58	.61	.63	.68	.67	.54	.51	.55
Cronbachs α utprövad uppgift	.873	.922	.869	.857	.819	.793	.809	.732	.912
Lösningsfrekvens ankaruppgift	.63	.65	.69	.59	.65	.62	.54	.62	.56
Cronbachs α ankaruppgift	.826	.806	.820	.825	.840	.836	.814	.778	.864

Tabell 2 visar att graden av uppskattning av de utprövade läsförståelseuppgifterna varierar från uppgift till uppgift, med medelvärden mellan 3.3 och 3.8[9]. Detta innebär att samtliga uppgifter som ingick i materialet fått förhållandevis positiva omdömen i genomsnitt; ingen fick ett medelvärde under 3. Som Tabell 2 visar har alla tre MC-uppgifter, en LL samt en OC fått de allra högsta omdömena, 3.7 eller 3.8. Hur bra en utprövningsuppgift upplevs vara beror naturligtvis inte enbart på uppgiftsformatet utan troligtvis i hög grad också på innehållet i texten. Elevernas omdömen kan vara en helhetsbedömning av såväl innehåll som format men det är också möjligt att somliga kan ha bedömt innehållet snarare än formatet, medan andra kan ha gjort tvärtom.

Vad gäller uppgifternas svårighetsgrad innebär ett lågt medelvärde i Tabell 2 att uppgiften ansetts svårare än när medelvärdet är högre. Tabell 2 visar att medelvärdet för de utprövade uppgifterna varierar mellan 2.4 och 3.2[10]. Fem medelvärden ligger under 3 vilket innebär att dessa uppgifter i genomsnitt uppfattats som förhållandevis svåra. Alla tre OC-uppgifter har medelvärden under 3, vilket också gäller för två LL. De tre uppgifterna av MC-typ upplevdes som lättare, med medelvärden mellan 3.1 och 3.2. I uppgifter av OC-typ måste eleven själv komma på det ord som saknas i luckan medan det i MC finns alternativ där eleven ska välja det ord som bäst passar in. Eftersom receptiv förmåga brukar föregå produktiv – man kan exempelvis oftast känna igen och förstå ord och fraser innan man själv kan använda dem i egen produktion – är det inte förvånande att elever upplever OC som något svårare än MC, eftersom OC har

[9] Standardavvikelse mellan 1.0 och 1.2.
[10] Standardavvikelse mellan 1.0 och 1.2.

inslag av produktion (jfr t.ex. Elgort & Nation, 2010). Det bör dock påpekas att orden i såväl MC som OC ofta är vanligt förekommande ord eftersom det är förmågan att läsa som avses prövas, inte ordförrådets omfång.

Även vad gäller hur bra eleverna upplevde att de klarade den utprövade uppgiften betyder ett högre medelvärde i Tabell 2 att de tror sig ha klarat uppgiften bättre än vid lägre medelvärde. Tabell 2 visar att medelvärdena varierar mellan 3.1 och 3.5[11] vilket indikerar att utprövningsgrupperna i genomsnitt upplevde att de klarade uppgifterna relativt bra eftersom ingen uppgift hamnade under 3. En LL och två MC-uppgifter upplevde eleverna att de klarade allra bäst med medelvärden på 3.4 eller 3.5. De uppgifter som eleverna i genomsnitt upplevde att de klarade sämst var en LL och två OC där medelvärdet var 3.1. Tabell 2 visar att de uppgifter som ansågs bäst också är de uppgifter som grupperna upplevde att de klarat bäst, nämligen LL1 och MC2 och MC3. De uppgifterna ansågs heller inte vara särskilt svåra.

Lösningsfrekvensen, det vill säga den andel items som respektive utprövningsgrupp i genomsnitt klarat, varierar mellan de utprövade uppgifterna. Resultaten bör tolkas i förhållande till gruppens lösningsfrekvens för ankaruppgiften eftersom denna uppgift är densamma i alla utprövningar. Tabell 2 visar att lösningsfrekvensen för ankaruppgiften varierar mellan .54 och .69[12] vilket indikerar att utprövningsgrupperna är olika starka. Grupperna som prövade ut MC1, OC1 och OC3 höll en något lägre nivå än övriga grupper att döma av lösningsfrekvensen på ankaruppgiften som var under .6 i dessa grupper och över detta värde i de andra. När en utprövad uppgift har en lägre lösningsfrekvens än ankaruppgiften kan det indikera att den har en högre svårighetsnivå än ankaruppgiften, och tvärtom om den har en högre lösningsfrekvens. Av de nio utprövade uppgifterna har fyra en högre lösningsfrekvens än ankaruppgiften, en LL och de tre MC-uppgifterna. Eleverna har alltså i genomsnitt klarat dessa fyra uppgifter bättre än ankaruppgiften. En OC har eleverna klarat lika bra som ankaruppgiften. Vad gäller övriga uppgifter har eleverna i genomsnitt något lägre resultat i de utprövade uppgifterna än i ankaruppgiften.

Reliabilitetsmåttet Cronbachs alpha varierar mellan .732 och .922 i de nio utprövade uppgifterna. Alla utom två uppgifter, MC 3 och OC 2 har alpha-värde över .8. I ankaruppgiften varierar värdet mellan .778 och .864.

Analys av samvariation

För att undersöka i vilken grad resultat på ankaruppgiften och resultat på den utprövade uppgiften samvarierade genomfördes en korrelationsanalys (Pearson). I analysen ingick samtliga resultat på ankaruppgiften och de nio utprövade uppgifterna. Resultatet visar att samvariationen är relativt stark och statistiskt signifikant (r = .73**). Det innebär att det i stor utsträckning är samma elever

[11] Standardavvikelse mellan 1.1 och 1.2

[12] Standardavvikelse mellan .2 och .3 i ankaruppgiften och de utprövade uppgifterna.

som har höga respektive låga resultat på både ankaruppgiften och utprövnings-uppgiften.

För att undersöka i vilken grad elevernas omdömen och uppfattningar om en uppgift samvarierade med det faktiska utfallet, det vill säga hur bra de klarade uppgiften när den poängsattes, genomfördes statistiska korrelationsanalyser (Pearson). Samvariationen mellan resultat på utprövningsuppgiften (lösnings-frekvens) och elevernas uppfattningar på en femgradig skala om hur bra upp-giften var, hur svår den var samt hur bra de upplevde att de klarade uppgiften analyserades. Resultat och enkätsvar från alla nio uppgifter ingick i analysen. Analysen visar att samvariationen mellan samtliga variabler är statistiskt signifi-kant. Samvariationen mellan elevers uppfattning om hur bra uppgiften var och lösningsfrekvensen är statistiskt signifikant men relativt låg (r = .38**), vilket också är fallet vad gäller samvariationen mellan upplevd svårighetsgrad och lös-ningsfrekvens (r = .34**). Samvariationen mellan hur bra eleven upplevde sig klara uppgiften och den faktiska lösningsfrekvensen är något högre (r = .47**).

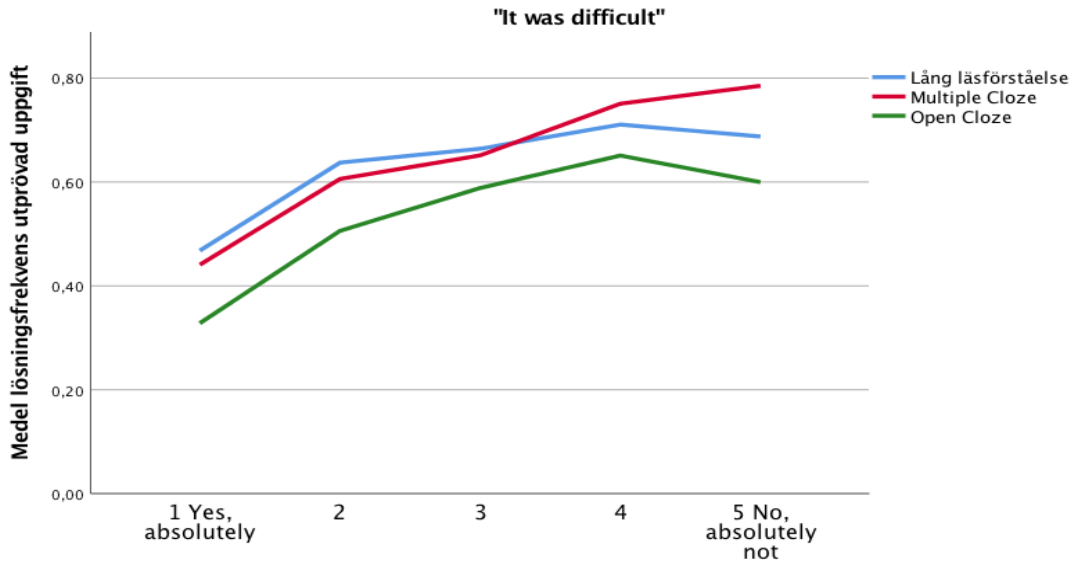
Vidare visar korrelationsanalysen att det finns en statistiskt signifikant samvariation mellan uppfattningen om provet, det vill säga hur bra eleven ansåg att provet var, och hur de upplevde att de klarat uppgiften (r = .54**). Resultatet kan tyda på att elever är benägna att uppleva att de klarar uppgifter bra om de tycker uppgiften är bra, eller att de anser att uppgiften är bra eftersom de klarar den bra.

Jämförelse mellan uppgiftstyper

Nedan följer tre figurer som visar samvariationen mellan elevernas svar på den femgradiga Likertskalan och lösningsfrekvensen för uppgiftstyperna LL, MC och OC. Som nämnts ingick tre uppgifter av varje uppgiftstyp.

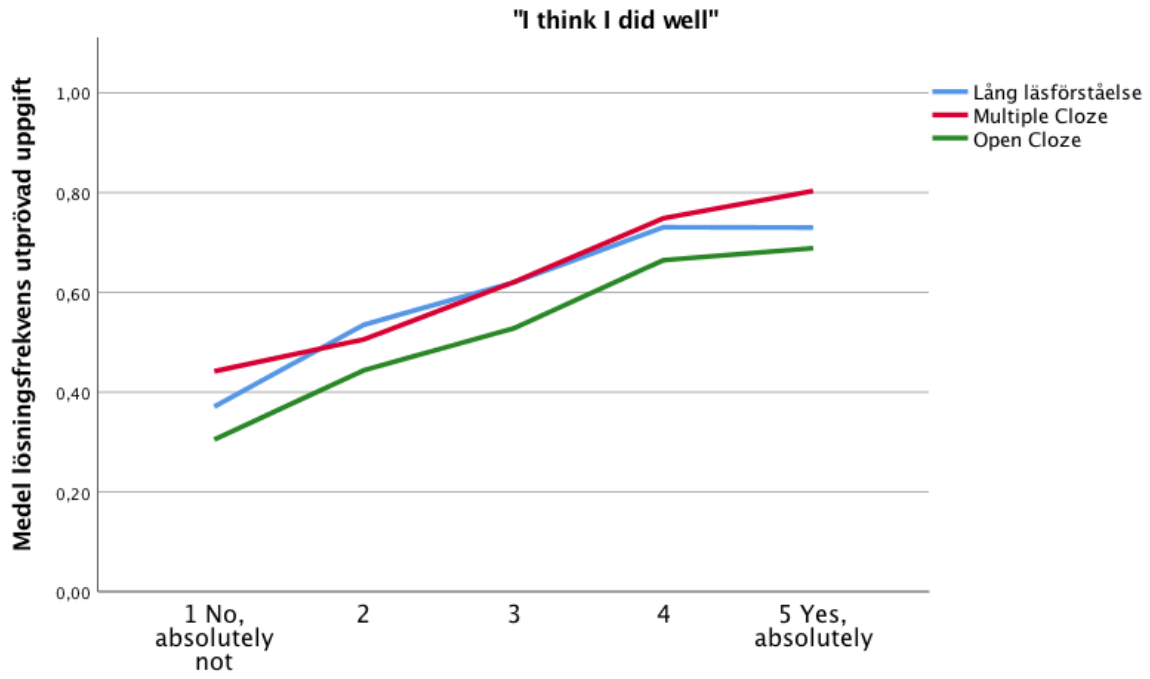
Figur 2. Lösningsfrekvens i förhållande till elevernas uppfattning om hur bra uppgiften var

Figur 2 visar att elever som gav uppgiften ett lågt omdöme i allmänhet klarade den sämre än elever som gav den högt omdöme. Det framgår också av Tabell 2, att den genomsnittliga lösningsfrekvensen för de OC som ingår i studien är lägre än för LL och MC.



Figur 3. Lösningsfrekvens i förhållande till elevernas uppfattning om uppgiftens svårighetsgrad

I Figur 3 framgår att den grupp elever som i enkäten markerat att uppgiften var svår uppnår en lägre lösningsfrekvens än de som markerat att den inte var det. Dock visar Figur 3 också att de elever som markerat att de utprövade OC- respektive LL- uppgifterna absolut inte var svåra, det vill säga 5 på Likertskalan, i genomsnitt klarat uppgiften något sämre än de elever som markerat 4.



Figur 4. Lösningsfrekvens i förhållande till elevernas uppfattning om hur bra de klarat uppgiften

Vad gäller elevernas uppfattningar om hur bra de upplevde att de klarade den utprövade uppgiften visar Figur 4 att elevernas uppfattningar verkar spegla utfallet som lösningsfrekvensen visar; ju säkrare man känt sig på att man klarat uppgiften bra, desto högre lösningsfrekvens.

Detaljerad analys av en uppgift

För att närmare analysera och illustrera i vilken grad elevers uppfattningar om en uppgift samvarierar med det faktiska utfallet visas här resultat av mer detaljerad analys av en uppgift, LL1. Korrelationen (Pearson) mellan lösningsfrekvens, det vill säga den andel av frågorna som eleverna klarade, och elevernas uppfattningar om uppgiften är statistiskt signifikant vad gäller intrycket av hur bra uppgiften var (r = .42**), hur svår den var (r = .39**) och hur bra de uppfattade att de klarade den (r = .45**). Korrelationsanalysen visar att det finns en viss samvariation mellan elevers uppfattningar om uppgiften och i vilken utsträckning de klarat den, vilket även framkom i Figurerna 2–4. Korrelationsanalysen visar också att elevernas uppfattningar inte alltid stämmer överens med hur väl de faktiskt lyckas.

I Tabellerna 3 och 4 visas lösningsfrekvensen för LL1 bland elever som markerat olika alternativ på den femgradiga skalan vad gäller uppgiftens svårighetsgrad respektive hur bra de ansåg sig ha klarat den.

Tabell 3. LL1 Lösningsfrekvens i förhållande till upplevd svårighetsgrad

	It was difficult				
	Yes, absolutely				No, absolutely not
	1	2	3	4	5
	N=39	N=65	N=138	N=99	N=34
Lösningsfrekvens LL1	.49	.56	.67	.74	.81

Tabell 4. LL1 Lösningsfrekvens i förhållande till hur väl eleven tror sig ha klarat uppgiften

	I think I did well				
	No, absolutely not				Yes, absolutely
	1	2	3	4	5
	N=23	N=41	N=119	N=131	N=60
Lösningsfrekvens LL1	.43	.48	.61	.74	.78

Tabell 3 och Tabell 4 visar att de elever som angett att uppgiften var svår klarade färre frågor än de elever som markerat att de inte tyckte uppgiften var svår. Likaså kan utläsas att de elever som angett att de klarat uppgiften bra, också klarade fler items än de som markerat att de inte klarat uppgiften bra.

En uppdelning mellan pojkar och flickor visar att den genomsnittliga lösningsfrekvensen på denna uppgift för pojkarna var .64 och för flickorna .68.

I Tabell 5 visas lösningsfrekvensen för pojkar respektive flickor som markerat olika alternativ på den femgradiga skalan vad gäller hur väl de upplevde att de klarade uppgiften.

Tabell 5. LL1 Jämförelse pojkar / flickor – lösningsfrekvens och självskattning

	I think I did well									
	No, absolutely not								Yes, absolutely	
	1		2		3		4		5	
Pojkar / flickor	p	f	p	f	p	f	p	f	p	f
Lösningsfrekvens LL1	.41	.44	.42	.54	.59	.62	.71	.78	.76	.82

Tabell 5 visar att lösningsfrekvensen bland flickor i genomsnitt är något högre än bland pojkar i varje steg av den femgradiga skalan. Man kan exempelvis se att de flickor som markerade en fyra på den femgradiga skalan – de upplevde att de klarade uppgiften bra men inte så bra att de markerade en femma – klarade uppgiften något bättre än de pojkar som markerat en femma, det vill säga att de absolut tyckte att de klarade uppgiften bra. Det tycks alltså som om pojkarna i genomsnitt har ett något bättre självförtroende än flickorna i detta avseende. En annan tolkning skulle kunna vara att pojkarna inte har lika höga krav på sig själva som flickorna.

Diskussion

Syftet med föreliggande studie var att undersöka och belysa på vilket sätt och i vilken omfattning test-taker feedback kan bidra till de nationella provens reliabilitet och validitet. I detta syfte analyserades test-taker feedback utifrån tre frågor som förekommit i samband med utprövning av uppgifter till det nationella provet i engelska för årskurs 9 under många år, nämligen frågor om hur bra respektive svår deltagarna ansåg att uppgiften var samt hur de upplevde att de klarade den. Resultaten visade att det fanns statistiskt signifikanta samvariationer mellan elevers uppfattningar och deras resultat när uppgifterna poängsattes, det vill säga ju bättre respektive lättare eleverna upplevde att uppgiften var, desto högre var deras resultat i genomsnitt. Likaså fanns en signifikant samvariation mellan hur bra eleverna upplevde att de klarade uppgiften och hur de faktiskt klarade den. Korrelationerna var dock relativt begränsade ($r < .5^{**}$). Det bör dock noteras att nio olika uppgifter av tre typer ingick i analysen och att utprövning av varje uppgift genomförts bland 400 elever per uppgift; alltså inte samma elever för alla uppgifter, vilket kan ha påverkat resultatet. Resultatet indikerar ändå att elevernas upplevelser av svårighetsgrad i viss utsträckning bekräftar det som deras resultat på uppgiften visar.

Det är givetvis så att elever upplever uppgifter på olika sätt, bland annat eftersom deras kunskapsprofiler och intressen kan skilja sig åt. Likaså kan elevers självkänsla skilja sig åt liksom de krav de ställer på sig själva, vilket också indikerades av studiens resultat. Exempelvis var flickornas skattning av hur bra de klarade en uppgift lägre än pojkarnas, med hänsyn tagen till det faktiska resultatet. För övrigt framkom liknande skillnader mellan flickor och pojkar i undersökningar redan under 1990-talet (Erickson, 1999). I föreliggande studie undersöktes inte orsakerna till sådana skillnader vidare, men i prov-utveckling kan denna typ av information föranleda att uppgiften granskas ytterligare för att undersöka om någon grupp gynnas eller missgynnas och vad det i så fall kan bero på. Test-taker feedback kan alltså, vilket exempelvis Brown (1993) visat, bidra till att stärka även den aspekt av validitet som relaterar till så

kallad *bias*, det vill säga om ett prov eller en uppgift gynnar eller missgynnar en särskild grupp på andra grunder än deras kunskapsnivå i engelska (se Kunnan, 2004).

Vad gäller elevernas uppfattningar om de tre uppgiftsformat som undersöktes i denna studie – lång läsförståelse, multiple cloze och open cloze – visade resultatet, som kunde förväntas, att eleverna tyckte det var svårare att själva fylla i ord i tomma luckor än att välja ett svar bland flera alternativ. Givetvis kan dock inga generella slutsatser dras av resultaten utifrån studiens begränsade empiri.

Analysen av test-taker feedback i föreliggande studie ger vid handen att det skulle kunna vara värdefullt att ställa ytterligare frågor eller mer ingående frågor till eleverna än de frågor som här använts eftersom det exempelvis inte alltid framgår vad det är som gör att en uppgift uppfattas som bra eller mindre bra, svår eller lätt. Man skulle exempelvis kunna separera frågor om innehåll och uppgiftsformat, eller ställa fler öppna frågor om positiva och negativa aspekter av uppgiften för att få mer detaljerad information.

Test-taker feedback kan ses som komplement till statistiska analyser av varje uppgifts och items validitet och reliabilitet (t.ex. Cronbachs alpha). Om exempelvis den statistiska analysen visar att en viss uppgift eller ett visst item har låg reliabilitet kan test-taker feedback ibland ge information som gör det möjligt att förtydliga eller på annat sätt förändra något i uppgiften, vilket kan stärka såväl reliabilitet som validitet. Informationen från test-taker feedback används vidare vid sammansättning och sekvensering av hela prov, det vill säga i vilken ordningsföljd olika uppgifter läggs in i provet eller ordningsföljden på items inom en viss uppgift. Härvid används informationen från eleverna, och upplevd svårighet jämförs med faktisk svårighet. Exempelvis läggs helst inte en uppgift som elever upplever som mycket svår först i ett prov även om det faktiska resultatet från utprövningen visat att de klarade uppgiften bra; detta för att inte försämra förutsättningar för elever att lyckas så bra som möjligt (jfr Erickson & Åberg, 2012; Erickson, 2010).

Ett nationellt provs främsta syfte är att vara ett stöd för lärares bedömning av elevers kunskaper i förhållande till de i kursplanen uppsatta kunskapskraven och därför måste provet innehålla uppgifter som möjliggör differentiering mellan olika betygsnivåer. Detta innebär att vissa elever kommer att uppleva en uppgift som svår medan andra uppfattar den som lätt. Det är mänskligt att tycka sämre om något som inte verkade gå så bra i provsammanhang. Ur validitetssynpunkt är det dock viktigt att undersöka om elever som uppnår låga resultat ändå verkar ha försökt lösa uppgifterna eftersom provet annars inte kan ge en rättvisande bild av dessa elevers kunskaper. Om man vid utprövningen ser att elever med låga resultat verkar ha gett upp, till exempel om de låtit bli att svara alls på en rad frågor, kan feedback ibland ge vägledning om varför, så att uppgifter kan modifieras. Även elever som har mycket goda kunskaper kan uppleva att uppgifter inte ger dem möjlighet att visa allt de kan om uppgifterna uppfattas som mycket lätta. Med det provsystem vi har nu, där elever i årskurs 9 prövas i

relation till kunskapskraven i årskurs 9 men inte mot högre stegs krav, är det svårt att tillgodose dessa elevers synpunkter. I en framtid skulle möjligen adaptiva prov kunna vara ett alternativ för att möta och bedöma elever på olika kunskapsnivåer.

Provutveckling kan beskrivas som ett stort grupparbete eftersom det är nödvändigt att ett stort antal elever och deras lärare samtycker till att genomföra utprövning; de statistiska beräkningarna av uppgifters och items validitet och reliabilitet bör grundas på ett relativt stort antal genomförda uppgifter. Att samtidigt fråga om deltagarnas synpunkter kan, som analysen visat, tillföra värdefull information (se även t.ex. Shohamy, 2001; Bachman & Palmer, 2010; Ryan, 2014). De flesta som gått eller arbetat i skolan kan säkert känna igen provsituationer där elever känt att de inte riktigt kommit till sin rätt – de kände inte att de fick möjlighet att visa vad de faktiskt kunde. Ett viktigt syfte med test-taker feedback är att ta reda på om elever upplever att provuppgifter ger dem möjlighet att visa vad de kan i engelska eftersom syftet med provet är just att pröva detta. Att involvera elever i provutvecklingen kan därför stärka provets giltighet och acceptans i den grupp som provet berör, det vill säga elever och deras lärare[13], vilket kan bidra till provets validitet (Messick, 1989; Little & Erickson, 2015).

Från hösten 2018 ska resultat på nationella prov *särskilt beaktas* vid betygssättningen, enligt Skolförordningen (SFS 2017:1107; se även Skolverket, 2018a). Även tidigare betonades att nationella provens viktigaste syfte var att ge stöd för en likvärdig bedömning men det behövde inte nödvändigtvis betyda att resultaten särskilt beaktades. Lärare kunde, och kan fortfarande, vid särskilda skäl bortse från nationella provresultat vid betygssättningen. Mot denna bakgrund är det intressant att se att de nationella provbetygen i engelska sedan länge har en hög grad av överensstämmelse med de betyg lärarna sätter på sina elever i ämnet – i allmänhet en högre grad av överensstämmelse än i andra ämnen i årskurs 9 där nationella prov ges (Gustafsson, Cliffordson & Erickson, 2014; Skolverket, 2018b). Detta ger en indikation om att lärares samlade bedömning av elevers kunskaper i engelska återspeglas tämligen väl i det nationella provet, trots att ett prov inte kan mäta alla aspekter av ett ämnes innehåll och krav. En bidragande orsak till den höga överensstämmelsen mellan provbetyg och betyg kan vara att utprövnings- och valideringsprocessen är omfattande och att den involverar ett stort antal elever och lärare, vilket troligtvis bidrar till en hög grad av acceptans för provet.

[13] Synpunkter samlas in även från undervisande lärare i de klasser som genomför utprövning. Det låg dock utanför denna studies ram att belysa dem.

Tack

Tack till anonyma reviewers och till redaktörerna för värdefulla kommentarer som hjälpte oss att utveckla texten. Tack också till Dorte Velling Pedersen, Gudrun Erickson och Jan-Eric Gustafsson för synpunkter och råd i olika faser av skrivandet, och till Maria Hulthén för noggrann korrekturläsning.

Om författarna

Eva Olsson är lektor i pedagogik vid Göteborgs universitet. Hennes forskningsintressen omfattar språkdidaktik och bedömning kopplat till elevers utveckling av vokabulär och förmåga att skriva, främst på engelska.
Institutionstillhörighet: Institutionen för pedagogik och specialpedagogik, Göteborgs universitet. Box 300, 405 30 Göteborg, Sverige
E-post: eva.olsson@ped.gu.se

Sofia Nilson är ansvarig för det nationella provet Engelska 6 för gymnasieskolan vid Göteborgs universitet.
Instituttionstillhörighet: Institutionen för pedagogik och specialpedagogik, Göteborgs universitet, Box 300, 405 30 Göteborg, Sverige
E-post: sofia.nilsson@ped.gu.se

AnnaKarin Lindqvist är projektledare för NAFS (Nationella prov i främmande språk) vid Göteborgs universitet.
Institutionstillhörighet: Institutionen för pedagogik och specialpedagogik, Göteborgs universitet, Box 300, 405 30 Göteborg, Sverige
E-post: annakarin.lindqvist@ped.gu.se

Referenser

Alderson, C., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. (1985). Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. *TESOL Quarterly, 19* (3), 535–556.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74.

Black, P. J. & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice.* DOI: 10.1080/0969594X.2018.1441807

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277–303.

Börjesson, L. (2012). *Om strategier i engelska och moderna språk*. Stockholm: Skolverket. http://www.skolverket.se/publikationer?id=3120

Chapelle, C. & Abraham, R. (1990). Cloze method: what difference does it make? *Language Testing*, *7*(2), 121–146. https://doi.org/10.1177/026553229000700201

Cheng, L. & DeLuca, C. (2011). Voices from test-takers: further evidence for language assessment validation and use. *Educational Assessment*, *16*(2),104–122. DOI: 10.1080/10627197.2011.584042

Cohen, A. D. (2007). The coming of age for research on test-taking strategies. I J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, E. & C. Doe, C (red.) *Language Testing Reconsidered* (s. 89–111). Ottawa: University of Ottawa Press.

Cronbach, L. J. (1984). *Essentials of psychological testing (4th Edition).* New York: Harper & Row

Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly, 1*(1), 5–18, DOI: 10.1207/s15434311laq0101_2

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*, 347–368.

Elgort, I., & Nation, I.S.P. (2010). Vocabulary learning in a second language: Familiar answers to new questions. I P. Seedhouse, S. Walsh & C. Jenks (red.), *Conceptualizing learning in applied linguistics* (s. 89–104). Houndmills: Macmillan.

Erickson, G. (1999). Från Sp 8 till Äp 9. Om utvecklingen av ett nytt nationellt prov i engelska i grundskolan. I I. Carlsson & N. H. af Ekenstam (red.), *Papers on Language. Learning, teaching, and assessment: Festschrift in honor of Torsten Lindblad*, (IPD Report no. 1999:02), (s. 202–230). Göteborg: Göteborgs universitet, Institutionen för pedagogik och didaktik.

Erickson, G. (2010). Good practice in language testing and assessment – a matter of responsibility and respect. I T. Kao & Y. Lin (red.), *A New Look at Teaching and Testing: English as Subject and Vehicle* (s. 237–258). Taipei, Taiwan: Bookman Books Ltd.

Erickson, G. (2016). Bedömningens dubbla funktion – för lärande och likvärdighet. *Communicare* (Fremmedspråksenteret, Norge) *6*, 4–9

Erickson, G. & Åberg-Bengtsson, L. (2012). A Collaborative Approach to National Test Development. I D. Tsagari & I. Czepes (red.), *Collaboration in Language Testing and Assessment* (s. 93–108). Frankfurt: Peter Lang Verlag

Fox, J. & Cheng, L. (2015). Walk a mile in my shoes: stakeholder accounts of testing experience with a computer-administered test, *TESL Canada Journal*, *32*, 65–86.

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics, 29*, 3–20. doi:10.1017/S0267190509090023

Grabe, W. (2009). *Reading in a second language: moving from theory to practice*. Cambridge: Cambridge University Press.

Grabe, W. & Stoller, F.L. (2011). *Teaching and researching reading*. (2nd ed.) Harlow, England: Longman/Pearson.

Gustafsson, J. E., Cliffordson, C. & Erickson, G. (2014) *Likvärdig kunskapsbedömning i och av den svenska skolan – problem och möjligheter.* Stockholm: SNS Förlag.

Hattie, J., & Timperley, H. (2007). The power of feedback, *Review of Educational Research, 77*, 81–112.

Hirsh, Å. & Lindberg, V. (2015). *Formativ bedömning på 2000-talet – en översikt av svensk och internationell forskning.* Stockholm: Vetenskapsrådet. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-120536

Huang, H.T. & Hung, S.T. (2017). EFL test-takers' feedback on integrated speaking assessment, *TESOL Quarterly*, *51*(1), 166–179. https://doi-org.ezproxy.ub.gu.se/10.1002/tesq.330

Huhta, A., Kalaja, P., & Pitkänen-Huhta, A. (2006). The discursive construction of a high-stakes test: The many faces of a test-taker, *Language Testing*, *23*(3), 326–350.

Kane, M. T. (2002). Validating high-stakes testing programs, *Educational Measurement: Issues and Practices*, *21*(1), 31–41.

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study, *Assessment in Education: Principles, Policy & Practice*, *22*(3), 302–323. DOI: 10.1080/0969594X.2014.988121

Klapp, A., Cliffordson, C. & Gustafsson, J. E. (2014). The effect of being graded on later achievement: evidence from 13-year olds in Swedish compulsory school, *Educational Psychology*, *36* (10), 1771–1789. DOI: 10.1080/01443410.2014.933176

Korp, H. (2011). *Kunskapsbedömning: Vad, hur och varför?* Stockholm: Skolverket

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation, *Language Testing*, *11*(3), 225–250

Kunnan, A. J. (2000). *Fairness and validation in language assessment*, Cambridge: Cambridge University Press.

Kunnan, A. J. (2004). Test Fairness. I M. Milanovic & C. Weir (red.), *European language testing in a global context* (s. 27–48). Cambridge: Cambridge University Press.

Little, D., & Erickson, G. (2015). Learner identity, learner agency, and the assessment of language proficiency: Some reflections prompted by the common European framework of reference for languages, *Annual Review of Applied Linguistics*, *35*, 120–139. doi:http://dx.doi.org/10.1017/S0267190514000300

Launeanu M., Hubley A.M. (2017). A model building approach to examining response processes as a source of validity evidence for self-report items and measures. I B. Zumbo & A. Hubley (red.) *Understanding and investigating response processes: Advances in validation research* (s. 115–136). New York, NY: Springer.

McCray, G. & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking, *Language Testing 35* (1), 51–73.

Messick, S. (1987). Validity, *ETS Research Report Series, 1987*(2), 1–208.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment, *Educational Researcher*, *18*(2), 5–11.

Moss. P. A. (2007). Reconstructing validity, *Educational Researcher*, *36*, 470–476.

Nevo, B. (1985). Face validity revisited, *Journal of Educational Measurement*, *22*(4), 287–293.

Nilholm, C. (2005). Specialpedagogik: Vilka är de grundläggande perspektiven? *Pedagogisk forskning i Sverige 10*, 124–138.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: a structural equation modelling approach*. Cambridge: Cambridge University Press.

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*, 441–474.

Ryan, D. E. (2014). Consider the candidate. Using test-taker feedback to enhance quality and validity in language testing, *e-TEALS: An e-journal of Teacher Education and Applied Language Studies, 5* : 1–23.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach, *Language Testing, 17*, 8–114.
Secolsky, C. (1987). On the direct measurement of face validity: A comment on Nevo, *Journal of Educational Measurement, 24* (1), 82–83
Shohamy, E. (2017). Critical language testing. I E. Shohamy, I. G. Or, & S. May (red.), *Encyclopedia of Language and Education* (3rd ed., pp. 441–454). Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-319-02326-7
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman/Pearson Education.
SFS 2017:1107. Förordning om ändring i skolförordningen (2011:185). Stockholm: Utbildningsdepartementet. Hämtad 2018-09-17, från https://www.lagboken.se/Lagboken/sfs/sfs/2017/1100-1199/d_3115971-sfs-2017_1107-forordning-om-andring-i-skolforordningen-2011_185
Skolverket (1994). *Läroplan för det obligatoriska skolväsendet, förskoleklassen och fritidshemmet*. Stockholm: Skolverket.
Skolverket (2011). *Läroplan för grundskolan, förskoleklassen och fritidshemmet*. Stockholm: Skolverket. https://www.skolverket.se/undervisning/grundskolan/laroplan-och-kursplaner-for-grundskolan
Skolverket (2018a). *Allmänna råd om betyg och betygssättning*. Stockholm: Skolverket. https://www.skolverket.se/publikationer?id=4000
Skolverket (2018b). *Relationen mellan provresultat och betyg i grundskolans årskurs 6 och årskurs 9 2017*. PM Enheten för förskole- och grundskolestatistik. https://www.skolverket.se/publikationer?id=3898
Spolsky B. (2017). History of language testing. I E. Shohamy, I. Or & S. May (red.) *Language Testing and Assessment*. *Encyclopedia of Language and Education* (3rd ed.), (s. 375–384).Cham: Springer International Publishing. https://doi-org.ezproxy.ub.gu.se/10.1007/978-3-319-02261-1_32
Stricker, L. J. (2012). *Testing: It's not just psychometrics* (ETS Research Memorandum No. RM- 12-07). Princeton, NJ: Educational Testing Service.
Taras, M. (2005). Assessment—summative and formative—some theoretical reflections, *British Journal of Educational Studies, 53*(4), 466–478.
Velling Pedersen, D. (2009). Trying to get it right. I *Språk och lärande. Rapport från ASLA:s höstsymposium, Stockholm, 7–8 november, 2008.* Stockholm: Association Suédoise de Linguistique Appliquée (ASLA).
Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing, 11*, 324–348. doi:10.1080/15305058.2011.589018
Zieky, M. (2006). Fairness review in assessment. I S. Downing & T. Haladyna (red.), *Handbook of Test Development* (s. 359–376). Mahwah, NJ: Lawrence Erlbaum.