Utbildning & Demokrati 2017, vol 26, nr 2, 113–137
Tema: Perspektiv på nationella prov

Hur tillförlitligt är det nationella provet i läsning i åk 9?

Michael Tengberg & Gustaf B Skar

Reliability of the national reading test in 9th grade. This article reports from a study of reliability in the Swedish national reading test in ninth grade. Based on a representative sample of 500 students, the study investigates internal consistency and power to discriminate between different levels of reading ability. Data was analyzed using both classical test theory and Rasch modelling. Results from the study suggest that neither the subscales nor the whole scale of items included in the test reach desirable levels of internal consistency. Three out of four subscales include items with strikingly low degree of internal correlation (.23–.46). Furthermore, results from the Rasch analysis suggest that the test is capable of reliable separation between two groups of test-takers only: lower performing students and higher performing students. The empirical support for categorizing student performances into six separate grade levels is therefore questionable. Implications for future development of the national test of reading are discussed.

Keywords: construct validity, national test, Rasch analysis, reading, reliability.

Inledning

Den här artikeln handlar om tillförlitlighet i det läsprov som ingår i det nationella provet i svenska i årskurs nio. Ämnesprovet i svenska består sammanlagt av tre delar: ett skrivprov, ett läsprov och ett muntligt prov. Alla tre är obligatoriska och alla tre bidrar till det sammanvägda provresultatet. I undersökningen tillämpas statistiska metoder för att

Michael Tengberg är docent i pedagogiskt arbete vid Karlstads universitet, institutionen för pedagogiska studier, 651 88 Karlstad. E-post: michael.tengberg@kau.se
Gustaf B Skar är projektledare och førsteamanuensis vid Skrivesenteret och Norges teknisk-naturvitenskapelige universitet i Trondheim, Nasjonalt senter for skriveopplæring og skriveforsking, 7491 Trondheim. E-post: gustaf.b.skar@ntnu.no

fastställa i vilken mån läsprovet lever upp till etablerad internationell standard för reliabilitet i kunskapsprov av motsvarande karaktär.

Diskussionen om de nationella proven i Sverige har under de senaste åren blivit alltmer intensiv. I stor utsträckning har det handlat om själva förekomsten av proven. Kritik har exempelvis riktats mot att nationella prov ges i för många ämnen och årskurser liksom mot utökad administrationen och mer press på både elever och lärare att förbereda, genomföra och efterarbeta proven, något som ibland upplevs ta orimligt mycket tid från undervisningen (Löfgren, Löfgren & Pérez Prieto 2015, Olovsson 2015, SOU 2016:25). Att studera de nationella proven är viktigt eftersom dessa får såväl didaktiska som juridiska konsekvenser för skolan. Didaktiska konsekvenser genom att de bidrar till att definiera ett ämnes innehållsdimension och operationalisera hur ämneskunnande kan gestaltas. Juridiska konsekvenser genom att provresultat bildar en betydelsefull del av betygsunderlaget och därmed inverkar på elevers möjligheter till framtida utbildnings- och yrkesval.[1] Vidare har provresultaten en potentiellt omfattande betydelse vid skolors marknadsföring. När staten inför prov med så pass omfattande konsekvenser måste medborgarna kunna lita på att provens resultat är både vederhäftiga och användbara. Därför är det nödvändigt att provens innehåll och tekniska kvaliteter underställs vetenskapliga prövningar.

Den tidigare forskningen om nationella prov i Sverige har i viss mån berört kvaliteten på själva innehållet. Det gäller exempelvis provinnehållets förmåga att representera kursplaneinnehållet, det vill säga provens förmåga att mäta grad av måluppfyllelse inom ämnet. Inom svenskämnet har dessa aspekter av provens validitet bland annat studerats av Eric Borgström (2012) och Michael Tengberg (2014a, 2014b). Vidare har Borgström och Per Ledin (2014) undersökt bedömarsamstämmighet på skrivprovet i gymnasiet (Svenska B) och Tengberg och Gustaf B. Skar (2016) har undersökt bedömarsamstämmighet för läsprovet i svenska i årskurs nio. Däremot saknas undersökningar som svarar på frågan om huruvida de olika uppgifterna i läsprovet hänger ihop och på ett trovärdigt sätt bidrar till mätning av läsförmåga antingen som delkompetenser eller som en sammanhängande kompetens, det vill säga antingen uttryckt i form av poäng på delskalor eller uttryckt i ett totalpoäng på hela provet. Vid sidan av innehållsdimensionen är frågan om konsistens avgörande för att kunna fästa tilltro till ett sammanlagt provresultat (Messick 1996, Bachman & Palmer 2010). Frånvaro av konsistens innebär svårigheter att fästa innebörd vid givna resultat, till exempel att ett visst provpoäng motsvarar den kompetens som uttrycks i ett betygssteg. Syftet med föreliggande studie är därför att bidra till kunskapen om de nationella provens

(och specifikt läsprovets i svenska i årskurs nio) validitet genom att undersöka rimligheten i att använda det samlade urvalet av uppgifter i provet som underlag för att mäta antingen delar av läsförmågan och/eller för att ge ett samlat provbetyg.

Att mäta läsförmåga

Att mäta individers läsförmåga är komplicerat eftersom läsförmåga som fenomen inte alldeles enkelt låter sig definieras och eftersom eventuella testresultat kommer att bero av en rad olika faktorer (Snow et al. 2002, Sabatini, Albro & O'Reilly 2012). Olika läsprov ställer exempelvis olika höga krav på lexikalitet, läsflyt, språkligt-diskursiva färdigheter, inferensskapande och inte minst på bakgrundskunskaper från vilka läsaren kan skapa sammanhang för att kunna värdera språk och innehåll i texter (Alderson 2000, Francis, Fletcher, Catts, & Tomblin 2005, Keenan, Betjemann & Olsson 2008). Vilka underliggande dimensioner i form av delkompetenser som läsförmågan egentligen består av är dock långt ifrån givet. Förslag på dimensionering i termer av *delkonstrukt*[2], eller så kallade läsprocesser, har lämnats av många forskare (Davies 1968, Khalifa & Weir 2009, Langer 1995, McNamara & Magliano 2009, VanderVeen, Huff, Gierl, McNamara, Louwerse & Graesser 2007). Det råder dock fortfarande oenighet om huruvida sådana delkonstrukt av läsförmåga verkligen existerar, vare sig psykometriskt eller konceptuellt (Alderson 2000, Rupp 2012). Ett flertal studier, baserade på korrelationsmätningar och faktoranalyser har visat att det är svårt att dela in läsförmågan i tillförlitliga delkonstrukt (Meijer & van Gelderen 2002, Rost 1993, Carey, Gordon, Schedl, & Tang 1996, Song 2008, Spearritt 1972, van Steensel, Oostdam & van Gelderen 2012). Trots detta är det vanligt att provkonstruktörer utifrån uppgiftsurvalet skapar delskalor som ska representera olika läsprocesser. Så görs exempelvis i PIRLS-provet, PISA-provet, det amerikanska NAEP-provet i läsning liksom i det svenska nationella provet (IEA 2009, National Assessment Governing Board 2013, OECD 2009, Skolverket 2015). I några fall är syftet med detta i första hand att skapa en bred representation av konstruktet, utan att poäng rapporteras per delskala (se t ex PIRLS, NAEP och det norska nationella läsprovet). I andra prov får delskalorna en mer definitiv betydelse antingen genom att poäng rapporteras på delskalenivå (t ex PISA[3]) eller, som i det svenska nationella provet, genom att elevers slutresultat eller provbetyg blir beroende av hur väl de presterat inom de olika delskalorna. När dimensionaliteten i provet får stora konsekvenser för resultaten måste valideringen emellertid också kunna

visa att provkonstruktionen håller för högre ställda psykometriska krav (van Steensel et al. 2012, Rupp 2012).

Validering av läsprov kan bygga på en rad olika tekniker. För mått på tillförlitlighet rapporteras exempelvis intern konsistens av uppgiftsurval som helhet och, i förekommande fall, av delskalor (Foorman, Koon, Petscher, Mitchell & Truckenmiller 2015, Francis, Snow, August, Carlson, Miller & Iglesias 2006, Roe 2014). Dessa indikerar i vilken grad provet förmår ta ett sammanhängande mått på det avsedda konstruktet eller på de avsedda delkonstrukten. Mäter de olika frågorna, som antas pröva samma slags läsprocess, verkligen samma sak? På läsprov är resultat dessutom ofta relaterade till textval genom testtagares bakgrundskunskaper och motivation för olika ämnesinnehåll, varför så kallat lokalt beroende ofta mäts för delar av testet. Längre texter med flera frågor till varje text, vilket är standard i svenska NP, inverkar som regel negativt på testets reliabilitet. Å andra sidan kan det motiveras med hänvisning till ekologisk validitet, det vill säga den utsträckning i vilken testresultaten ger goda skattningar av naturligt förekommande läsning både i skolan och i övriga livet. Vidare har Miyoko Kobayashi (2002) i en studie av ett läsförståelsetest för japanska andraspråksläsare visat att textstruktur och responsformat kan ha statistiskt signifikanta effekter på testtagares resultat. Mer välstrukturerade texter och öppna uppgifter bidrog till större differentiering mellan resultaten för hög- och lågpresterande läsare. För prövning av konstruktvaliditet och processvaliditet görs också olika former av kvalitativa analyser av uppgiftsurval (Campbell 2005, Tengberg 2014a) liksom flerdimensionella analyser där kvalitativa och kvantitativa ansatser vägs samman. Joseph Magliano, Keith Millis, Yasuhiro Ozuru & Danielle S. McNamara (2007) har t ex utvecklat ett ramverk för hur processer, strategier och texter för elever på olika färdighetsnivåer kan jämföras mellan olika former av bedömningsverktyg.

För analyser av huruvida testtagare presterar som förväntat på hela eller delar av ett givet uppgiftsbatteri liksom om uppgifterna selekterar mellan elever på olika färdighetsnivåer på ett förväntat sätt har det blivit vanligt med undersökningar baserade på Rasch-modellering (Rasch 1980, Bond & Fox 2015) och Item Response Theory (IRT) (Hohensinn & Kubinger 2011, Rauch & Hartig 2010). Rasch-analysen har en rad mättekniska fördelar, bland annat möjliggör den en estimering av testtagarnas färdighetsnivå och uppgifternas svårighetsgrad på en gemensam intervallskala, den så kallade logit-skalan (McNamara 1996). Bland fördelarna med Rasch-analys brukar också nämnas att det är möjligt att erhålla elev- och item-estimat som är oberoende av varandra, något som är en förutsättning för att på ett tillförlitligt sätt skilja elever med olika förmåga och uppgifter med olika svårighet.

Användning av estimat från Rasch-analysen bygger på antagandet om psykometrisk endimensionalitet, det vill säga att mätning av ett eller flera delkonstrukt eller läsprocesser tillsammans ger konsistenta totalpoäng (jfr McNamara 1996). När en provuppgift således inte passar Rasch-modellen indikerar det att uppgiften mäter något annat.

Rasch-analysen bygger på den enkla modellen att en elev har 50 % chans att klara en uppgift, som är lika svår som eleven är duktig. Ju större diskrepans mellan elev och uppgifter, desto större respektive mindre blir sannolikheten att eleven svarar korrekt. För att ett visst dataset skall kunna sägas passa Rasch-modellen måste de empiriska observationerna ge stöd åt detta antagande. Om så inte är fallet kan estimat av elevers läsförmåga inte användas för att förutspå elevens framgång på olika uppgifter. I sin tur är det en stark indikation på att uppgifterna sammantaget inte mäter läsförmåga på ett konsistent sätt.

I den internationella forskningen om läs- och skrivprov är Rasch-analyser legio. Youn-Hee Kim och Eunice E. Jang (2009) har exempelvis använt DIF-analys (differential item functioning) för att visa hur olika grupper av testtagare (här första- respektive andraspråkselever) gynnas eller missgynnas av olika delkomponenter i ett provinsgemensamt literacytest i Ontario, Kanada. Rene T. Proyer, Michaela M. Wagner-Menghin och Gyöngyi Grafinger (2014) har använt Rasch-modellering för validering av uppgiftsurval i läsprov för vuxna. Dominique P. Rauch och Johannes Hartig (2010) har i en studie jämfört hur väl olika IRT-modeller (endimensionell och tvådimensionell) förmår analysera läsprovsresultat och med den tvådimensionella modellen bland annat visat att en väsentlig del av variansen av läsprestationer bara fångas av constructed response-uppgifter (CR), vilket bekräftar Kobayashis (2002) resultat. Christine Hohensinn och Klaus D. Kubinger (2010), som också använde IRT-modellering för att undersöka effekter av responsformat i språkfärdighetstest fann inga skillnader med avseende på vilken latent förmåga som mättes med olika format, men däremot att multiple choice-uppgifter (MC) var signifikant lättare än CR-uppgifter, något som Yo In'nami och Rie Koizumi (2009) tidigare påvisat även för rena läsförståelsetest.

Studiens problemställning och forskningsfrågor

Det nationella provet i svenska i årskurs nio spelar en betydande roll för elevers slutbetyg i ämnet. Proven har också betydelse på en mer övergripande nivå genom att resultatstatistik används som argument dels i olika former av politiskt och ekonomiskt beslutsfattande, dels

av föräldrar som väljer skolor åt sina barn. Eftersom provresultatens konsekvenser kan vara omfattande både för skolor och elever bör även bevisen för giltigheten hos de slutsatser som baserar sig på provresultaten vara av det mer omfattande slaget (Kane 2015). Det innebär exempelvis att vi vill kunna försäkra oss om att de olika provuppgifterna tar ett tillförlitligt mått på den aktuella populationens färdigheter och förmågor, liksom att vissa grupper av testtagare inte missgynnas av provets utformning. Givet detta kombinerar vi i denna studie klassisk testteori med Rasch-analyser (som båda beskrivs utförligare nedan) för att svara på följande frågeställningar:

1. I vilken utsträckning är delskalorna i provet respektive skalan som helhet konsistenta?
2. I vilken utsträckning förmår läsprovet skilja mellan olika nivåer av läsförmåga?

Metod

Urval

Urvalet består av elevlösningar från 500 slumpvist valda elever i nionde klass som genomförde det nationella provet i svenska (läsdelen) vid ordinarie provtillfälle i mars 2015 under 60 min (inläsning) + 140 min provtid. Materialet har samlats in av Nationella provgruppen i Uppsala. Totalt genomfördes provet i svenska av 82 255 elever samt i svenska som andraspråk av 9 353 elever.

I urvalet ingår både elever som läst svenska (N = 450) och elever som läst svenska som andraspråk (N = 50).[4] Med tillgänglig statistik över provresultat som publiceras på Skolverkets hemsida (http://www.skolverket.se/statistik-och-utvardering/provresultat/) kan vi se hur eleverna i urvalet presterar i jämförelse med elever i riket. Av elever som läser svenska är andelen som på delprov B (läsa) fått lägst betyget E 90,7 % i urvalet och 90,0 % i riket. För elever som läser svenska som andraspråk är motsvarande siffror 72,0 % i urvalet respektive 58,4 % i riket. Hur eleverna i urvalet respektive i riket fördelar sig över betygsskalan redovisas i tabell 1.

Tabell 1. Fördelning av delprovsbetyg i urvalet och i riket.

		F	E	D	C	B	A	Poäng[a]
Svenska	Urval	9,3%	15,6%	18,4%	32,4%	13,1%	11,1%	13,2
	Riket	10,0%	16,2%	17,3%	32,8%	12,4%	11,2%	13,1
Svenska 2	Urval	28,0%	18,0%	22,0%	20,0%	12,0%	0,0%	9,7
	Riket	41,6%	22,3%	14,7%	16,8%	3,4%	1,2%	7,4

[a] Genomsnittligt provbetygspoäng

Sammantaget ser vi att bland elever som läser svenska presterar eleverna i urvalet för studien marginellt bättre än riksgenomsnittet. Någon signifikant skillnad i betygsfördelning mellan urval och riksgenomsnitt finns det dock inte, $\chi^2(5, N = 82\ 624) = 0{,}85$, $p = 0{,}97$. När det gäller elever som läser svenska som andra språk är skillnaden något större till fördel för eleverna i urvalet. Här är skillnaden med avseende på betygsfördelning dessutom statistiskt signifikant, $\chi^2(5, N = 9\ 403) = 16{,}00$, $p = 0{,}01$. Därmed kan vi alltså inte dra några generella slutsatser om provets egenskaper specifikt för elever som läser svenska som andraspråk.[5]

Provet

Samtliga delprov i det nationella provet i svenska utvecklas vid Institutionen för nordiska språk vid Uppsala universitet på uppdrag av Skolverket. Nationella prov avser mäta den enskilde elevens grad av måluppfyllelse i enlighet med kunskapskraven i respektive kursplan. Därigenom, menar uppdragsgivaren, syftar proven till att dels stödja en likvärdig och rättvis bedömning och betygsättning, dels ge underlag för analys av måluppfyllelse på huvudmannanivå, kommunnivå och nationell nivå (Skolverket 2014). Mer specifikt syftar läsprovet i årskurs nio till att pröva elevens förmåga att "läsa och analysera skönlitteratur och andra texter för olika syften" (Skolverket 2014, s 20).

Provet består av 20 uppgifter som är kopplade till sju texter av varierande längd (från ca 100 ord till ca 2000 ord). Ungefär hälften av textmaterialet består av skönlitteratur och andra hälften av sakprosa. Uppgifterna har för avsikt att pröva fyra olika läsförståelseprocesser (*hitta efterfrågad information*; *dra enkla slutsatser*; *sammanföra och tolka information och idéer samt reflektera*; och *granska och värdera innehåll, språk och textuella drag*). Dessa är i stort sett identiska med de läsprocesskategorier som används i PIRLS-provet (IEA 2009) och konceptuellt mycket nära besläktade med de tre läsförståelseprocesser som prövas i PISA (OECD 2009) liksom i det norska nationella läsprovet (Tengberg 2017). Uppgifterna består vidare av såväl MC-uppgifter som

CR-uppgifter (se nedan). I tabell 2 redovisas fördelning av uppgifter per läsförståelseprocess och responsformat.

Tabell 2. Uppgifter fördelade på läsprocess och responsformat.

Läsförståelseprocess	Antal MC	Antal CR	Antal totalt
Hitta efterfrågad information	1	1	2
Dra enkla slutsatser	1	3	4
Sammanföra och tolka	2	8	10
Granska och värdera	2	2	4
Totalt	6	14	20

MC-uppgifterna är av standardformat med ett korrekt alternativ och tre distraktorer.[6] CR-uppgifterna består av uppgifter som kräver såväl längre som kortare svar. Ibland räcker det med ett ord. För andra uppgifter får testtagarna upp till 10 rader att formulera svar på. Bedömningsskalorna är tillika av olika skallängd och ger 0–2 p, 0–2–4 p, 0–2–4–6 p eller 0–2–4–6–8 p. I internationell jämförelse kännetecknas det svenska nationella läsprovet av att det totala antalet uppgifter i provet är få samt att en stor del av dem är CR-uppgifter med långa bedömningsskalor (se t ex Tengberg 2017 för en jämförelse med norska och danska nationella läsprov).

Analys

För att besvara den första forskningsfrågan har vi använt klassisk mätteoretisk analys av intern konsistens för såväl delskalor med avseende på läsprocess som för provet som helhet. Intern konsistens refererar till den grad i vilken olika uppgifter (items) mäter samma underliggande konstrukt. Här har vi använt Cronbach's alpha (för lättillgänglig beskrivning, se Bachman 2004) och Spearman-Brown-korrelationer. I normalfallet estimeras bara alpha, men för skalor som har färre än tre uppgifter rekommenderas i första hand att Spearman-Brown används istället (Eisinga, te Grotenhuis & Pelzer 2012). Alpha bygger på genomsnittet av korrelationen mellan de olika uppgifterna i provet och antalet uppgifter. Uppgifterna om riktvärden för alpha vid storskaliga tester varierar något, delvis på teoretiska grunder, men även med hänsyn tagen till vad som testas och hur testresultaten ska användas (Kaplan & Saccuzzo 2009, Nunnally 1975). För prov som är av stor betydelse för testtagarna kan värden under .70 ofta anses svaga, värden runt .80 som acceptabla och värden omkring .90 och högre betraktas som tecken på god intern konsistens (Kaplan & Saccuzzo 2009, Kline 2000). För läsprov specifikt saknas såvitt vi kunnat finna allmänt etablerade gränsvärden. Det är emellertid värt att nämna att

reliabilitetskoefficienter i PISA 2009 för svenska data låg på .91–.92 både för lässkalan som helhet och för de tre delaspekterna (OECD 2012). Likaså anger tekniska rapporter från norska nationella läsprov alphavärden på omkring .90 (se t ex Eriksen & Roe 2013, Roe 2014).

För att svara på den andra forskningsfrågan har vi genomfört Rasch-analyser. I sin enklaste form ser Rasch-modellen ut såhär:

$$Ln(Pni1/Pni0) = Bn - Di,$$

där *Pni1* är sannolikheten att elev *n* svarar rätt på uppgift *i* och *Pni0* är sannolikheten att samma elev svarar fel på denna uppgift.[7] B*n* är elevens skattade förmåga (personestimat) och D*i* är uppgift i:s skattade svårighetsgrad (uppgiftsestimat). Ett första steg i Rasch-analysen är att utgå från antalet uppgifter som en elev svarat rätt på och antalet elever som klarat en viss uppgift. Dessa resultat ligger till grund för person- och uppgiftsestimat. Genom analysen blir dessa estimat invarianta, det vill säga oberoende av varandra. En uppgifts svårighet förändras inte beroende på vilka personer som löser den och personers förmåga förändras inte beroende på vilka uppgifter de möter. Skulle modellen visa att så inte är fallet beror det typiskt på att en eller flera uppgifter inte mäter samma konstrukt som övriga uppgifter (Engelhard 2013).

Såväl elevens förmåga som uppgiftens svårighetsgrad uttrycks som logit-värden på den så kallade logit-skalan (logit = log-odds unit). Denna antar i vanliga fall värden mellan -5,0 och 5,0. Uppgifters genomsnittssvårighet antar av konvention värdet 0,00. För att kunna placera både elever och uppgifter på samma skala genomförs en icke-linjär logistisk transformering, från ordinal- till intervall-skala (Engelhard 2013). Detta i sin tur innebär att avstånden på skalan bli ekvidistanta. Exempelvis blir det då lika långt mellan 0 och 1, som mellan 1 och 2. Det betyder att vi på ett trovärdigt sätt kan skatta avståndet mellan elever med olika förmågor och mellan uppgifter av olika svårighetsgrad.

Rasch-analysen är lämplig i krävande situationer, så som den som präglar svenska NP. Här har vi att göra med 20 uppgifter, varav hälften är dikotoma och hälften polytoma, det vill säga att eleven kan få grader av korrekt svar. Dessutom opererar man i NP med tre olika skallängder, vilka redovisats ovan. I programmet *Winsteps* 3.92.1 (Linacre 2016) har en Rasch-analymodell som tar hänsyn till detta implementerats:

$$Ln(Pnij/ Pni(j-1)) = Bn - Dgi - Fgj$$

Den nya termen, *Fgj*, avser punkten på logit-skalan där skalsteg j är lika sannolik som skalsteg j-1. Bokstaven ”g” avser den grupp uppgifter (här: grupp g) som delar skallängd.

Rasch-analysen bibringar en rad användbara mått. I den här undersökningen skall vi fokusera på följande:

- Variabel-kartan, som illustrerar hur väl uppgifternas svårighetsgrad matchar elevernas förmågenivåer.
- Separationsstatistiken, som anger i vilken utsträckning provet lyckats särskilja elevgruppen i olika förmågenivåer.
- Kategoristatistiken, som redovisar frekvensen elever för givna nivåer och genomsnittlig förmåga förbunden med dessa nivåer, vilket ger indikationer på hur väl skalorna till de olika uppgifterna fungerar.

För att kunna genomföra Rasch-analysen i *Winsteps* har vi re-kodat materialet genom att ta bort de “tomma” skalstegen och således exempelvis låtit 0-2 bli 0-1 och 0-2-4 till 0-1-2. Detta innebär att den tidigare totala poängskalan från 0–66 ersatts med en ny, från 0–33.

Som inledande analys har vi också undersökt den så kallade model fit-statistiken, som för varje uppgift (och varje person) anger i vilken utsträckning data passar modellen. Detta anges i termer *infit* och *outfit*. När data passar modellen perfekt är detta värde 1,0. När data är för oförutsägbar överstiger värdet 1,0. Det kan exempelvis handla om provuppgifter med oförutsägbara egenskaper. När värdet är lägre än 1,0 är data redundanta, det vill säga de tillför inte någon mer information antingen om provets egenskaper eller om testtagarnas förmåga. Riktlinjerna för hur mycket större värdet kan vara varierar, men vi ansluter oss till John M. Linacre (2016), som anger 0,50–1,50 som intervall för fungerande uppgifter. Visar sig en eller flera uppgifter inte passa modellen kan det finnas fog för re-analys utan denna eller dessa uppgifter. I föreliggande material låg alla uppgifter innanför detta intervall (se bilaga 1).[8]

Resultat

Intern konsistens

Intern konsistens för provet som helhet samt för var och en av de fyra delskalorna anges i tabell 3 nedan. Som framgår är konsistensen sett till ovan angivna referensvärden låg både för provet som helhet och för

de olika delskalorna. I fall där det råder hög konsistens inom delskalorna, men där dessa inte korrelerar särskilt högt med varandra kan det förklara en lägre konsistens för testet som helhet. I det här fallet är emellertid även alphavärden (liksom Spearman-Brown för delskala 2) för delskalorna påfallande låga, med undantag för delskala tre (Sammanföra och tolka) för vilken korrelationen åtminstone närmar sig en acceptabel nivå. För övriga delskalor är den interna konsistensen så låg att det inte finns grund för att hävda att uppgifterna inom respektive processkategori mäter något sammanhängande delkonstrukt. Antalet uppgifter är visserligen få, men elevernas prestation på de olika uppgifterna korrelerar också i alltför liten grad med varandra. Analysen visar att det finns uppgifter som endast korrelerar .15 (uppg. 4) respektive .06 (uppg. 8 och 20) med övriga uppgifter inom respektive delskala. Trots att antalet uppgifter inom delskalorna är få skulle den interna konsistensen inom delskalorna öka om dessa uppgifter togs bort. Sammanfattningsvis kan vi konstatera att föreliggande provresultat inte utgör någon tillförlitlig grund för slutsatser om elevers färdigheter på läsprocessnivå. Inte heller blir det rimligt att låta resultat inom läsprocesskategorierna få betydelse för provbetyget.

Tabell 3. Intern konsistens.

Skala	Antal uppgifter	Cronbach's α	Spearman-Brown korr.
Hela provet	20	.77	
1. Hitta information	2	.22	.29
2. Dra enkla slutsatser	4	.46	
3. Sammanföra och tolka	10	.68	
4. Granska och värdera	4	.23	

Den här delen av analysen rör alltså i första hand delskalorna på provet. När det gäller rimligheten i att använda det samlade provresultatet som grund för slutsatser om elevernas läsförmåga ska vi använda oss av Rasch-analysen. Detta blir också logiskt mot bakgrund av att delskalorna visat sig inte vara konsistenta och provet därmed inte ska ses som bestående av fyra separata delkonstrukt. Konsistensanalysen ger snarare fog för att betrakta uppgifterna i provet som delar av ett sammanhängande konstrukt, något som också utgör förutsättningen för att kunna tillgripa Rasch-analys (jfr McNamara 1996).

Rasch-analysen

Rasch-analysen gav ett medelvärde bland eleverna på 0,98 logits (S.D. 0,99) och vi ser i figur 1 att många elever ligger tätt samlade kring ±

1 logit. Rasch-analysen visar vidare att elevernas förmåga inte tycks matcha uppgifternas svårighet särskilt väl. I figur 1, variabel-kartan, ser vi att ett antal elever har en förmåga som vida överstiger den svåraste uppgiftens svårighet. I mitten av figuren ser vi uppgifter där svårighetsgraden motsvarar elevförmågan, det vill säga uppgifter som elever på motsatt sida strecket har 50 % chans att klara.

Uppgifterna är ordnade efter svårighetsgrad och vi ser att uppgift 20 (U20) är den svåraste och uppgift 1 (U1) är den lättaste. Studerar vi logit-värdena ser vi att det skiljer 5,17 logits mellan U1 (-3,18) och U20 (1,99). Att uppgifterna inte samlat sig fullt så koncentrerat som figur 1 antyder har med U1 att göra. Bara 2 % av eleverna svarar fel på den. Bland eleverna är spridningen med avseende på förmåga större och här skiljer det 6,56 logits mellan den med lägst resultat och den med högst (-2,9 för elev #204 och 3,66 för elev #495). Konverterat tillbaka till poängskalan på provet motsvarar det senare 2 respektive 32 poäng (eller 4 respektive 64), det vill säga nästan alla fel eller nästan alla rätt.

Variabel-kartan indikerar att det finns flera uppgifter som inte bidrar med särskilt mycket information om elevernas läsförmåga. Majoriteten av eleverna har en förmågenivå som överstiger uppgiftssvårigheten hos nära hälften av uppgifterna i provet (U1, U2, U5, U7, U8, U10, U15, U18 och U 19). Det betyder att eleverna har mycket hög chans att klara dem. Givet det begränsande antal uppgifter som ingår i provet finns det därför endast ett fåtal uppgifter som på ett nyanserat sätt kan särskilja elever ovanför dessa nivåer.

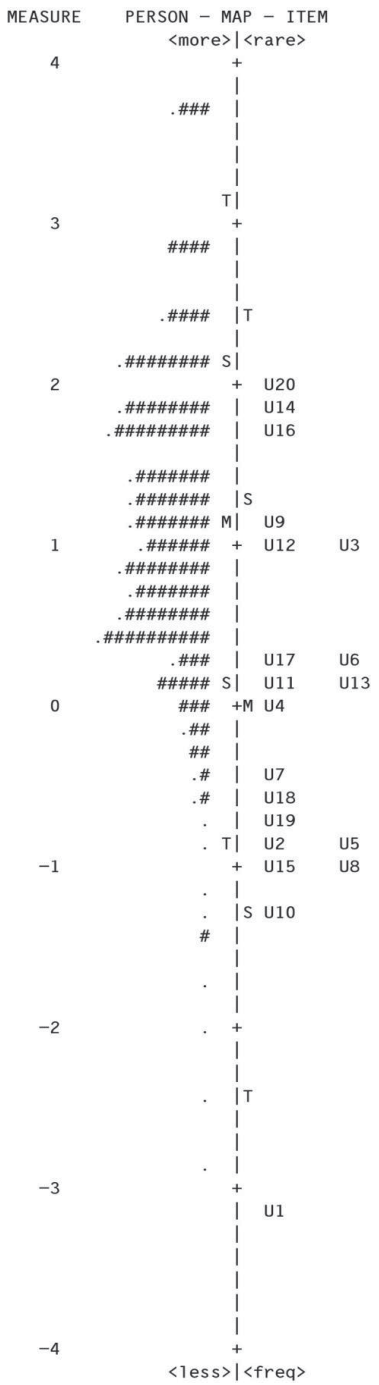
För att komplettera informationen från variabel-kartan presenterar vi i tabell 4 deskriptiv kategoristatistik. Denna redovisar antalet förekomster av de olika kategorierna för varje uppgift, samt det genomsnittliga person-logitvärdet förbundet med dessa. Enligt Rasch-modellen skall elever med högre förmåga alltid ha större sannolikhet att nå högre resultat, varför det finns anledning att vara observant på kategorier där detta inte stämmer. En sådan inkongruens kan indikera att det är något fel med bedömningsguiden eller med svarsalternativen.

Som vi kan se i tabellen följer alla uppgifter principen om att elever med högre logit-värde får högre resultat. Vi ser också att, vilket är att förvänta, att eleverna i varierande grad fått rätt snarare än fel på uppgifterna. En extrem uppgift är U2, där 489 elever fått rätt och 11 fått fel. U20 och U9 är exempel på uppgifter som bryter mot mönstret: U9 genom att vara en svår uppgift som majoriteten elever inte klarar och U20 genom att vara en uppgift där grupperna i respektive kategori är ungefär lika stora. För de polytoma uppgifterna är det generella mönstret att en kategori dominerar, det vill säga en stor majoritet av eleverna hamnar på samma skalsteg i bedömningen, ofta det högsta skalsteget. Därefter dominerar någon av mellankategorierna, vilket

gör kategori 0 till en sällan använd kategori. Ett undantag här är U12, som har kategori 0 och kategori 4 som största kategorier.

Separationsstatistiken, slutligen, kan användas för att skatta i hur många distinkta förmågenivåer ett prov lyckas särskilja elever. Utgångspunkten är genomsnittliga mätfel för hela provet (för tekniska detaljer, se t ex Schumacker & Smith 2007). *Winsteps* rapporterar olika mått på separation, varav vi rapporterar *R*, som anger med vilken reliabilitet ett test har skiljt mellan olika nivåer av förmåga och ett separationsindex, *G*, som anger hur många grupper av förmågenivåer testet urskiljer. *R* kan maximalt anta värdet 1,0, medan *G* inte har någon sådan begränsning. Linacre (2016) anger att *R*-värden runt 0,9 indikerar 3 till 4 nivåer av förmåga, runt 0,8 indikerar 2 eller 3 nivåer och runt 0,5 runt 1 eller 2 nivåer.

Resultatet i den här undersökningen visar en personseparation på $G = 1{,}78$ och $R = 0{,}76$ och en uppgiftsseparation på $G = 10{,}18$, $R = 0{,}99$.[9] Medan uppgifternas svårighetsgrad alltså kalibreras med hög precision, lyckas testet inte tydligt särskilja ens mellan två distinkta nivåer av elevförmåga. Med andra ord indikerar resultaten att provet i bästa fall ger stöd för att placera elever i endera av två olika kategorier av läsförmåga (jfr Linacre 2016).



Figur 1. Variabel-kartan. Längst till vänster syns logit-skalan, från -4,0 till 4,0. I nästa kolumn, "personer" representeras elever av punkter (.) och staket (#), beroende på antal. I kolumnen längst till höger återfinns uppgifterna. Eleverna är ordnade så att elever med hög förmåga har höga logit-värden och vice versa. Uppgifterna som är relativt sett svåra har höga logit-värden och uppgifter som relativt sett är enkla har låga logit-värden.

Tabell 4. Deskriptiv kategoristatistik

Uppgift	Kategori	Observationer	Andel %	Logit medel	Logit S.D.	Logit S.E.
U1	0	11	2	-0,20	0,74	0,23
	1	489	98	1,13	0,98	0,04
U2	0	17	3	-0,50	1,02	0,25
	1	88	18	0,60	0,94	0,10
	2	395	79	1,28	0,90	0,05
U3	0	241	48	0,70	0,79	0,05
	1	259	52	1,47	1,02	0,06
U4	0	144	29	0,60	0,96	0,08
	1	356	71	1,31	0,93	0,05
U5	0	39	8	-0,06	1,13	0,18
	1	53	11	0,68	0,90	0,12
	2	408	82	1,27	0,89	0,04
U6	0	99	20	0,22	0,92	0,09
	1	129	26	0,97	0,78	0,07
	2	272	54	1,48	0,88	0,05
U7	0	42	8	-0,14	0,97	0,15
	1	105	21	0,87	0,85	0,08
	2	353	71	1,32	0,91	0,05
U8	0	73	15	0,42	1,03	0,12
	1	427	85	1,22	0,94	0,05
U9	0	255	51	0,69	0,89	0,06
	1	245	49	1,53	0,91	0,06
U10	0	60	12	0,10	1,01	0,13
	1	440	88	1,24	0,91	0,04
U11	0	82	16	0,17	0,89	0,10
	1	129	26	0,70	0,73	0,06
	2	289	58	1,55	0,85	0,05
U12	0	142	28	0,28	0,72	0,06
	1	38	8	0,60	0,71	0,12
	2	94	19	0,90	0,68	0,07
	3	59	12	1,21	0,55	0,07
	4	167	33	1,99	0,75	0,06
U13	0	41	8	-0,26	0,91	0,14
	1	77	15	0,40	0,71	0,08
	2	153	31	0,96	0,66	0,05
	3	229	46	1,67	0,84	0,06
U14	0	331	66	0,89	0,94	0,05
	1	169	34	1,52	0,96	0,07
U15	0	73	15	0,35	1,16	0,14
	1	427	85	1,23	0,90	0,04
U16	0	234	47	0,59	0,86	0,06
	1	178	36	1,26	0,70	0,05
	2	88	18	2,14	0,91	0,10
U17	0	81	16	0,21	0,97	0,11
	1	160	32	0,88	0,73	0,06
	2	259	52	1,52	0,91	0,06
U18	0	47	9	-0,19	0,94	0,14
	1	64	13	0,54	0,73	0,09
	2	389	78	1,35	0,87	0,04
U19	0	87	17	0,21	1,01	0,11
	1	413	83	1,29	0,88	0,04
U20	0	339	68	0,94	0,94	0,05
	1	161	32	1,45	1,01	0,08

Notera. Kategori: skalsteg/poäng; Observationer: antalet elever som tilldelats visst poäng; Andel %: andelen elever som tilldelats visst poäng; Logit: genomsnittligt logit-värde (för personer) förbundet med kategorien; Logit S.D.: standardavvikelsen för kategorien; Logit S.E.: standardfelet.

Diskussion

Denna undersökning har fokuserat två kvalitetsaspekter av det svenska nationella läsprovet för elever i årskurs 9, nämligen delskalornas och den hela skalans interna konsistens samt provets möjlighet att skilja mellan olika nivåer av läsförmåga. Med andra ord har undersökningen behandlat två grundläggande aspekter av provets reliabilitet. Att reliabiliteten är hög är nödvändigt, om än inte tillräckligt, för att testresultat ska kunna läggas till grund för valida slutsatser om elevers läsning (jfr Kane 2015).

Skalanalysen har visat att provets delskalor inte bärs upp av uppgifter med stark korrelation till varandra. Därmed ger provresultaten inte någon tillförlitlig grund för slutsatser om elevers förmåga på läsprocessnivå. Vidare är konsistensen för provet som helhet i underkant av de nivåer som kan förväntas av ett prov med så pass höga insatser för både skolor och elever. Rasch-analysen har visat att elevresultaten efter transformeringen till linjär skala dels klumpar ihop sig, dels är förbundna med så stora mätfel att provet med säkerhet knappt kan urskilja 2 grupper utifrån förmågenivå: en grupp med ”lågpresterande” elever och en med ”högpresterande”.

Givet den omfattande forskningen på området är de låga konsistensvärdena på delskalorna inte särskilt överraskande. I flera andra standardiserade prov undviker man att ta mått på delskalor just eftersom det är erkänt svårt att erhålla höga nivåer av intern konsistens när antalet uppgifter per delskala är för få (OECD 2009). Delkonstrukt i form av läsprocessnivåer är dessutom inte bara svåra att belägga empiriskt, vilket redogjorts för ovan, utan även svåra att fastställa konceptuellt (DeStefano, Pearson & Afflerbach 1997, Tengberg accepterad). Oaktat antalet uppgifter i testet måste uppgifterna ifråga förstås också representera de konceptuella skillnader som delkonstrukten antas markeras av. Frågan är således om det är delkonstrukten som behöver omdefinieras eller om det är operationaliseringen av dem som fordrar bättre konstruktion och utprövning av uppgifter.

Oavsett vilket innebär den låga konsistensen problem. Eftersom uppgifterna som operationaliserar delkonstrukten inte hänger samman är det oklart om två olika råpoäng på en delskala (t ex ”granska och värdera”) faktiskt representerar två olika nivåer av förmåga på den delskalan. Det innebär i sin tur att trovärdigheten i resultat på delskalenivå faller och att provbetyg baserade på delresultat inte är att rekommendera.

Rasch-analysen har blottlagt en skillnad mellan å ena sidan användningen av råpoäng och å andra sidan det statistiska stödet för en sådan användning. Att det saknas fog för att dela upp elevernas resultat i sex distinkta nivåer av läsförmåga kan förstås tolkas som goda nyheter; i

ett kriterierelaterat prov förväntar vi oss inte en given fördelning och efter adekvat undervisning är det inte säkert att det föreligger någon större spridning bland elevresultaten. Men eftersom provresultaten används som om det finns sex nivåer är det viktigt att se på andra tänkbara förklaringar till att Rasch-analysens transformering till linjär skala inte kan bekräfta denna användning.

I litteraturen (t ex Bachman 2004, Bond & Fox 2015, Linacre 2016) anges fyra vanliga orsaker till att ett prov inte förmår skilja mellan elevförmåga, nämligen:

- faktisk spridning av elevförmåga (spridning i elevurvalet),
- testlängden,
- antal skalsteg per uppgift och
- matchning mellan elevförmåga och uppgiftssvårighet.

I det svenska nationella provet är elevurvalet givet, men testlängd, skalsteg och matchning är åtgärdbara. Vad gäller testlängd kan konstateras att provet, allt annat lika, skulle behöva utökas med hela 100 uppgifter för att uppnå en reliabilitet på R = ,95.[10] En sådan våldsam ökning av antalet uppgifter är förstås inaktuell. Därför behöver man sannolikt vidta en kombination av åtgärder för att öka reliabiliteten. Dessa åtgärder inkluderar fler uppgifter, fler svårare uppgifter och kanske förändrade bedömningsregler. Exempelvis noterades tidigare en knapp användningen av skalsteget 0. Möjligen beror den på att tröskeln för att få delvis rätt (2 poäng eller mer) är för låg och i ett prov där majoriteten av elever får rätt eller gradvis rätt på så få uppgifter som 20, finns begränsade möjlighet att med säkerhet göra distinktioner i flera läsförmågenivåer.

Sammanfattningsvis har den här studien dokumenterat bristande reliabilitet i det svenska nationella provet i läsning för elever i årskurs nio. Studien har också resulterat i ett antal frågor som, givet provets centrala betydelse, bör besvaras grundligt. I kommande undersökningar bör därför antagandet om existensen av delkonstrukt prövas, både kvantitativt och kvalitativt. Vidare bör kommande undersökningar studera något som inte varit möjligt här, nämligen hur olika distraktorer fungerar. Detta är ett viktigt inslag vid utvecklingen av uppgifter och vanligt förekommande vid utformningen av större läsprov (OECD 2012, Roe 2014).

Den här studien bygger på provresultat från 500 slumpvis valda elever från hela Sverige. Av den grunden finns anledning att tro att resultaten går att generalisera till testpopulationen som helhet. Med en sådan utgångspunkt blir studiens viktigaste bidrag att den kunnat dokumentera att användningen av råpoängsresultaten både som uttryck

för olika delkompetenser och som underlag för provbetyg tycks sakna statistiskt stöd. Detta är allvarligt givet de konsekvenser som användning av resultaten från nationella provet medför.

Noter

1. Nationella prov i Sverige är inte examensprov utan förväntas utgöra en del av lärarens samlade information om elevernas kunskaper.
2. Med *konstrukt* avses den specifika egenskap eller förmåga som ett test avser mäta. Delkonstrukt syftar således på eventuellt förekommande delaspekter av konstruktet.
3. Observera att detta sker i PISA endast de år då läsning utgör huvudområde för undersökningen och antalet uppgifter är fler. Övriga år, det vill säga då matematik eller naturvetenskap utgör huvudområden, anses antalet uppgifter till läsning vara för få för tillförlitlig rapportering på delskalenivå (OECD 2009).
4. Urvalet samlades in av Nationella provgruppen i Uppsala, som ansvarar för de nationella proven i svenska och svenska som andraspråk. I ämnet svenska har urvalet dragits från elever födda den 18 februari, 18 maj, 18 augusti och 18 november. I svenska som andraspråk har urvalet dragits från elever födda i maj. I båda fallen gäller att urvalet utgörs av de 450 respektive 50 först inkomna elevlösningarna.
5. För detta ändamål hade det varit nödvändigt med ett större och mer representativt urval. Föreliggande urval utgörs av samtliga elevlösningar som den nationella provgruppen samlade in 2015 och något större urval har vi inte haft tillgång till. Det dragna urvalet fungerar dock väl för att undersöka provets egenskaper för elever som läser Svenska specifikt liksom för att undersöka provegenskaper som inte kan antas bero av vilket av svenskämnena som eleverna läser.
6. En distraktoranalys (se t ex Tarrant, Ware & Mohammed 2009) kunde också ha bidragit till viktig information om provets och uppgifternas funktionalitet, men i den här studien har vi inte haft tillgång till rådata över elevlösningar, utan endast till elevresultat per uppgift. Korrekt svar på MC-uppgifter ger två poäng på provet.
7. Av utrymmesskäl presenterar vi de olika Rasch-modellenrna otekniskt och koncists. Vi avstår också från att fördjupa oss i de antaganden om lokalt oberoende (t ex att möjligheten att svara på en fråga bygger på att ha svarat rätt på en annan) och ”endimensionalitet” som är förbundna med Rasch-analyser. Den finns dock en mängd lättillgänglig litteratur för den intresserade läsaren (Bond & Fox 2015, Boone, Staver & Yale 2014, McNamara 1996, Linacre 2016).

8. Vår utgångspunkt är med andra ord att resultaten är tillräckligt sammanhängande för att Rasch-analysen skall vara meningsfull, något som också indikeras av Cronbach alpha-värdet för testet som helhet. Hade vi applicerat strängare krav (t ex intervallet 0,8-1,2 som används av PISA) hade denna utgångspunkt varit felaktig. En mer utförlig analys av stödet för detta antagande har uteslutits av utrymmesskäl.
9. För personer: Root mean square measurement error (RMSE): 0,48, observerad standardavvikelse (S.D.): 0,99, justerad S.D.: 0,87. För uppgifter, RMSE: 0,12, observerad S.D.: 1,26, justerad S.D. 1,22.
10. Detta antagande stöds av beräkning av Spearman-Brown Prophecy Formula, som utifrån beftinlig data kan användas för att beräkna hur många fler uppgifter ett prov behöver för en given reliabilitet. Spearman Brown Prophecy Formula återges i Linacre (2016) och ställs upp på följande sätt: T = C * RT * (1-RC) / ((1-RT) * RC), där T är antalet uppgifter; C är nuvarande antal uppgifter; RT är den önskade reliabiliteten; RC är den nuvarande personreliabiliteten.

Referenser

Alderson, J. Charles. (2000): *Assessing reading*. Cambridge: Cambridge University Press.

Bachman, Lyle F. & Palmer, Adrian S. (2010): *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, Lyle. F. (2004): *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bond, Trevor. G. & Fox, Christine. M. (2015): *Applying the Rasch Model*. New York: Routledge.

Boone, Wiliam J., Staver, John R. & Yale, Melissa S. (2014): *Rasch analysis in the human sciences*. New York: Springer.

Borgström, Eric (2012): Skrivförmåga på prov. I Gustaf Skar & Michael Tengberg red: *Svenskämnet i går, i dag, i morgon*, s 209–223. Stockholm: Svensklärarföreningen.

Borgström, Eric & Ledin, Per (2014): Bedömarvariation. Balansen mellan teknisk och hermeneutisk rationalitet vid bedömning av skrivprov. *Språk och Stil, 24*, 133–165.

Campbell, Jay R. (2005): Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. I Scott G. Paris, & Steven A. Stahl, red: *Children's Reading Comprehension and Assessment*, s 347–368. Mahwah, New Jersey: Lawrence Erlbaum Ass.

DeStefano, Lizanne, Pearson, P. David., & Afflerbach, Peter (1997): Content validation of the 1994 NAEP in Reading: Assessing the relationship between the 1994 assessment and the reading framework. I Robert Linn, Robert Glaser, & George Bohrnstedt, red: *Assessment in transition. 1994 trial state assessment report on reading: Background studies*, s 1–50. Stanford, CA: The National Academy of Education.

Eisinga, Rob, te Grotenhuis, Manfred & Pelzer, Ben (2012): The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? *International Journal of Public Health, 58*(4), 637–642.

Engelhard, George (2013): *Invariant Measurement.* New York: Routledge.

Eriksen, Anna & Roe, Astrid (2013): *Den nasjonale prøven i lesing på 8. og 9. trinn, 2013. Rapport basert på populasjonsdata.* Oslo: Institutt for lærerutdanning og skoleforskning. Universitetet i Oslo.

Foorman, Barbara R., Koon, Sharon, Petscher, Yaacov, Mitchell, Alison & Truckenmiller, Adrea (2015): Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology, 107*(3), 884–899.

Francis, David J., Fletcher, Jack M., Catts, Hugh & Tomblin, Bruce (2005): Dimensions affecting the assessment of reading comprehension. I Scott G. Paris & Steven A. Stahl, red: *Children's reading comprehension and assessment, s* 369–394. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Francis, David J., Snow, Catherine E., August, Dianne, Carlson, Coleen D., Miller, Jon & Iglesias, Aquiles (2006): Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension, *Scientific Studies of Reading, 10*(3), 301–322.

Hohensinn, Christine & Kubinger, Klaus D. (2011): Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732–746.

In'nami, Yo & Koizumi, Rie. (2009): A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*, 219–244.

IEA (2009): PIRLS 2011 Assessment framework. Chestnut Hill, MA: Boston College.

Kane, Michael T. (2015); Validitet. I G. Skar & M. Tengberg, red.: *Bedömning i svenskämnet*,s 212–237. Stockholm: Natur och kultur.

Kaplan, Robert M. & Saccuzzo, Dennis P. (2009): *Psychological testing: principles, applications, and issues*. (7. ed.) Belmont, Calif.: Wadsworth.

Keenan, Janice M., Betjemann, Rebecca S. & Olson, Richard K. (2008): Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281–300.

Khalifa, Hanan, & Weir, Cyril J. (2009): *Examining reading: Research and practice in assessing second language learning*. New York: Cambridge University Press.

Kim, Youn-Hee & Jang, Eunice E. (2009): Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning: A Journal of Research in Language Studies, 59*(4), 825–865.

Kline, Paul (2000): *Handbook of psychological testing*. London: Routledge.

Kobayashi, Miyoko (2002): Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*(2), 193–220.

Langer, Judith (1995): *Envisioning literature: Literary understanding and literature instruction*. New York & London: Teachers College Press.

Linacre, John M. (2016): *Winsteps® Rasch measurement computer program User's Guide*. Program manual 3.92.0. Beaverton, Oregon: Winsteps.com.

Löfgren, Håkan, Löfgren, Ragnhild & Pérez Prieto, Hector (2015): Preparations for the national tests in grade six – narratives from those that are assessed. Paper presenterat vid NERA-konferensen 2015, 3–5 mars, Göteborg.

Magliano, Joseph P., Millis, Keith, Ozuru, Yasuhiro & McNamara, Danielle S. (2007): A multidimensional framework to evaluate reading assessment tools. I Danielle S. McNamara,

red: *Reading strategies: Theory, interventions, and technology*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
McNamara, Tim. F. (1996): *Measuring Second Language Performance*. London: Longman.
McNamara, Danielle S. & Magliano, Joseph P. (2009): Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation, 51*, 297–384.
Meijer, Joost & Van Gelderen, Amos (2002): *Lezen voor het leven: Een empirische vergelijking van een nationale en een internationale leesvaardigheidspeiling*. Amsterdam: SCO-Kohnstamminstituut.
Messick, Samuel (1996): Validity and washback in language testing. *Language Testing 13*(3), 241–256.
National Assessment Governing Board (2013): *Reading framework for the 2013 National Assessment of Educational Progress*. Washington: U.S. Department of Education.
Nunnally, Jum C. (1975): *Introductory statistics for psychology and education*. New York: McGraw-Hill.
OECD (2009): *PISA 2009. Assessment framework: Key competencies in reading, mathematics and science*. Retrieved from http://www.oecd.org
OECD (2012): *PISA 2009 Technical Report*. OECD Publishing. Retrieved from http://www.oecd.org
Olovsson, Tord G. (2015): *Det kontrollera(n)de klassrummet. Bedömningsprocessen i svensk grundskolepraktik i relation till införandet av nationella skolreformer.* (diss.) Umeå: Umeå universitet.
Proyer, Rene T., Wagner-Menghin, Michaela M. & Grafinger, Gyöngyi (2014): Screening reading comprehension in adults: Development and initial evaluation of a reading comprehension measure. *Psychological Test and Assessment Modeling, 56*(4), 368-381.
Rasch, George (1980): *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.
Rausch, Dominique P. & Hartig, Johannes (2010): Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.
Roe, Astrid (2014): *Den nasjonale prøven i lesing på 8. og 9. trinn, 2014. Rapport basert på populasjonsdata*. Oslo: Institutt for lærerutdanning og skoleforskning. Universitetet i Oslo.
Rost, Detlef H. (1993): Assessing different components of reading comprehension: Fact or fiction? *Language Testing, 10*(1), 79–92.

Rupp, André A. (2012): Psychological vs. psychometric dimensionality in reading assessment. I John Sabatini, Elizabeth R. Albro, & Tenaha O'Reilly, red: *Measuring up: Advances in how we assess reading ability*, s 135–152. New York: Rowan & Littlefield Publishers, Inc.

Sabatini, John, Albro, Elizabeth & O'Reilly, Tenaha (2012): *Measuring up: advances in how we assess reading ability*. New York: Rowan & Littlefield Publishers, Inc.

Carey, Patricia A., Gordon, Ann, Schedl, Mary A. & Tang, K. Linda (1996): *An analysis of the dimensionality of TOEFL reading comprehension items*. Princeton, NJ: Educational Testing Service.

Schumacker, Randall E. & Smith, Everett V. (2007): Reliability. A Rasch Perspective. Educational and *Psychological Measurement, 67*(3), 394–409.

Skolverket (2014): *Lärarhäfte. Nationella provet i svenska/svenska som andraspråk 2014. Del A. Att läsa och förstå*. Stockholm: Skolverket.

Skolverket (2015): *Lärarhäfte. Nationella provet i svenska/svenska som andraspråk 2015. Del A. Att läsa och förstå*. Stockholm: Skolverket.

Snow, Catherine E. (red.) (2002): *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand.

Song, Min-Yong (2008): Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing, 25*(4), 435–464.

Spearritt, Donald (1972): Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*, 92–111.

Tarrant, Marie, Ware, James, & Mohammed, Ahmed M. (2009): An assessment of functioning and nonfunctioning distractors in multiple-choice questions: A descriptive analysis. BMC *British Medical Education, 9*, 40. http://dx.doi.org/10.1186/1472-6920-9-40

Tengberg, Michael (accepted): Validation of sub-constructs in reading comprehension tests using teachers' classification of cognitive targets. *Language Assessment Quarterly.*

Tengberg, Michael (2014a): Konstruktion och bedömning av förmågan att läsa (och förstå) skönlitterär text. Två nedslag i det nationella provets läsförståelsedel, åk 9. *Educare*, 2014:1, 79–97.

Tengberg, Michael (2014b): Att pröva elevers läsförmåga i årskurs nio. En analys av uppgiftkonstruktioner i det nationella provet

i svenska, delprov A ”Att läsa och förstå”. *Utbildning & Demokrati, 23*(2), 109–132.
Tengberg, Michael (2017): National reading tests in Denmark, Norway, and Sweden. A comparison of construct definitions, cognitive targets, and response formats. *Language Testing, 34*(1), 83–100.
Tengberg, Michael & Skar, Gustaf B. (2016): Samstämmighet i lärares bedömning av nationella prov i läsförståelse. *Nordic Journal of Literacy Research, 2*(1), 1–18.
Utbildningsdepartementet (2016): *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning.* SOU 2016:25. Utbildningsdepartementet: Stockholm.
VanderVeen, Arthur, Huff, Kristen, Gierl, Mark, McNamara, Danielle S., Louwerse, Max & Greasser, Art (2007): Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT Reasoning Test. I Danielle S. McNamara, red: *Reading comprehension strategies: Theories, interventions, and technologies*, s 137–172. New York: Lawrence Erlbaum Ass.
van Steensel, Roel, Oostdam, Ron & van Gelderen, Amos (2012): Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing* 30(1), 3–21.

Bilaga

Sammanfattande statistik för uppgifter

Uppgift	Logit	Obs.	Råpoäng	Logit S.E.	Infit	Intif_z	Outfit	Outfit_z
1	-3,18	500	489	0,31	1,02	0,18	0,56	-1,10

2	-0,92	500	878	0,1	0,95	-0,52	1,10	0,66
3	1,01	500	259	0,1	1,00	-0,07	1,10	1,89
4	0,03	500	356	0,11	1,05	0,95	1,04	0,50
5	-0,84	500	869	0,09	1,22	2,26	1,35	2,19
6	0,35	500	673	0,07	1,07	1,21	1,06	0,78
7	-0,39	500	811	0,08	1,07	0,89	1,15	1,25
8	-0,96	500	427	0,13	1,03	0,33	1,05	0,35
9	1,15	500	245	0,1	0,97	-0,81	0,97	-0,60
10	-1,22	500	440	0,15	0,92	-0,73	0,80	-1,16
11	0,19	500	707	0,07	0,95	-0,80	0,88	-1,35
12	0,99	500	1071	0,04	0,96	-0,56	1,17	1,63
13	0,16	500	1070	0,06	0,90	-1,62	0,86	-1,87
14	1,9	500	169	0,1	1,08	1,71	1,19	2,51
15	-0,96	500	427	0,13	0,97	-0,32	1,14	0,92
16	1,72	500	354	0,07	0,85	-2,93	0,82	-2,46
17	0,33	500	678	0,07	0,98	-0,39	1,01	0,17
18	-0,61	500	842	0,09	1,06	0,77	0,88	-0,94
19	-0,73	500	413	0,13	0,91	-1,25	0,84	-1,21
20	1,99	500	161	0,1	1,13	2,67	1,29	3,56
Medel	0,00	500	566,95	0,11	1,00	0,05	1,01	0,29

Notera. Logit: placering på logitskalan; Obs: antalet gånger uppgiften är tagen; Råpoäng: antalet poäng som delats ut; Logit S.E.: standardfelet för logit-värdet; Infit: infit-värden; Infit_z; t-statistik (signifikans vis <2 och >2); Outfit: outfit-värden; Outfit_z; t-statistik (signifikans vis <2 och >2):