MALMÖ UNIVERSITY | EDUCARE

2023:1

doi: 10.24834/educare.2023.1.698

Generering av provbetyg i nationella prov - Exemplet Religionskunskap årskurs 9

Frank Bach

https://orcid.org/0000-0001-5614-9064

frank.bach@gu.se

In the Swedish national tests, the so-called "Beläggsmodell" has been used to generate test grades in seven school subjects. This article examines the origin and effects of the "Beläggsmodell" on test scores in one of these subjects, religious studies. Classical methods for evaluating the quality of tests are used in combination with confirmatory factor analysis. The results show that the tests have a reliability above 0.92 and that the "Beläggsmodell" reduces this in unpredictable ways for individual students. Confirmatory factor analysis shows that the tests are mainly one-dimensional, which has the consequence that the "Beläggsmodell" does not fulfil any summative function. Instead, it has major consequences for individual students' test grades. There are examples of students who at the same total score on the test receive everything from D to A as a test grade. The large differences arise through differences in performance on a very small number of items. If, instead, the total sum is used as a basis for generating test grades, the possibilities for fair and equivalent test grades increase, as shown by examples. Responsible authorities should and, as announced, will switch to a system where the total sum points are the basis for the test grades.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

1. Inledning och litteraturgenomgång

År 2011 fick Sverige en ny läroplan för grundskolan med nya kursplaner som för första gången omfattade kunskapskrav i stället för betygskriterier. Parallellt med detta beslutades att det skulle införas nationella prov i geografi, historia, religionskunskap och samhällskunskap för årskurs nio utöver de som redan fanns i biologi, engelska, fysik, kemi, matematik och svenska (även svenska som andraspråk). Därutöver beslutades det också om nationella prov för årskurs sex i samma ämnen som i nian samt prov i årskurs 3 i matematik och svenska (och svenska som andraspråk). De första skarpa proven i SO-ämnena för årskurs nio genomfördes 2014 och det senaste genomfördes våren 2022. Proven har på grund av Corona-pandemin blivit inställda 2020 och 2021. Proven från våren 2022 ingår inte i denna studie.

1.1 Tidigare forskning

Olika aspekter av validitet kopplat till de första versionerna (2013-2015) av det nationella provet i religionskunskap i Sverige har studerats relativt väl. Främst problematiseras etik som innehåll i provet och hur väl det som mäts i provet stämmer med intentionerna i kursplanen och med forskning om etik och undervisning i etik samt vilka utmaningar och begränsningar som är förknippade med utvärdering av såväl elevers kunskaper i etik som deras handlingar (Franck, 2017; Osbeck m.fl., 2015; Osbeck m.fl., 2018). Stern (2017) argumenterar därutöver för en etik som rör själva bedömningen av elevers etiska kompetens genom prov medan Tykesson (2017) undersöker om uppgifterna i provet diskriminerar mellan olika grupper av elever på ett rättvist sätt. Tykesson visar att pojkar presterar relativt sett bättre på flervalsuppgifter medan flickor erhåller högre bedömning på uppgifter med öppna svar. När det gäller elever som har svenska som andra språk har de en fördel på uppgifter som fokuserar islam, men Tykesson noterar att gruppen med elever som har svenska som andra språk är i minsta laget för att dra säkra slutsatser (Tykesson, 2017). De två sistnämnda texterna knyter an till arbeten av Messick (1989a) och Kane (2016) som båda betonar etik när det gäller prov där de sist nämnda menar att oönskade negativa effekter talar emot ett testprograms validitet, då begreppet också inbegriper konsekvenser av tolkningar och beslut kopplade till provet. Fancourt (2017) lyfter frågan om att mycket små kvalitativa skillnader i elevsvar på enskilda uppgifter kan leda till stora skillnader i omdömen, allt från E till A,

trots att (de dåvarande) kunskapskravens formuleringar snarare pekade mot ett kontinuum i skalan. Han för ett resonemang om tänkbara orsaker men leder inte diskussionen explicit vidare till de validitetsfrågor som beror av reliabilitet i de provbetyg som genereras.

Forskning som modellerar provet i religionskunskap eller näraliggande ämnen har inte påträffats i litteratursökning inför arbetet med denna artikel. Däremot är det vanligare i andra skolämnen, exempelvis i matematik och engelska och inte minst i anslutning till storskaliga studier som PIRLS[1], TIMSS[2] och PISA[3]. Ett gemensamt drag för många av proven är att en generell, övergripande faktor har ett tydligt inflytande på alla komponenter i ett prov (Carlstedt & Gustafsson, 2005; Erickson m.fl., 2015), eller att urskiljbara faktorer är korrelerade (Brunner m.fl., 2012; se Figure 2 i Zhang m.fl., 2021). Även när ett provbatteri med många skolämnen modelleras finns betydande samvariation mellan olika delar (Carlstedt & Gustafsson, 2005; Åberg-Bengtsson & Erickson, 2006).

1.2 Bakgrund och studiens relevans

Syftet med de nationella proven har ändrats genom åren, från att i perioder endast ha ett summativt syfte att rättvist stödja lärare betygssättning till att i andra också stödja formativt inriktat arbete. Proven har också haft som syfte att utgöra ett redskap för implementering av kursplaner (Gustafsson & Erickson, 2018). Numera är syftet mer renodlat att stödja en rättvis betygssättning.

Provbetyg för det nationella provet i religionskunskap genereras med hjälp av en så kallad beläggsmodell. Med beläggsmodell avses att bedömningen av ett elevsvar på en enskild uppgift utgör ett belägg för en viss betygsnivå. En tidig beskrivning av beläggsmodellen finns i en skolverksrapport från 2010 om utprövningsproven i NO-ämnena 2009 (Bårman m.fl., 2010):

[1] https://www.iea.nl/studies/iea/pirls

[2] https://www.iea.nl/studies/iea/timss

[3] https://www.oecd.org/pisa/

Frank Bach

> För att åstadkomma en stark och tydlig koppling till kursplanens mål och kriterier, samt ge ett bra underlag för att visa på elevernas styrkor och svagheter, har en beläggsmodell tagits fram som används i bedömningen av elevernas svar. Beläggen fördelas i en matris som därmed ger en kunskapsprofil av vilka kvaliteter ur de tre aspekterna som eleven har samlat på det nationella ämnesprovet.
>
> Tanken är att elevens svar ska visa belägg för de specifika mål och kriterier som krävs för att besvara uppgiften. Bedömningen handlar om att avgöra om elevens svar ger belägg för kunskapen eller inte.

En grundläggande idé med beläggsmodellen är alltså att elevsvar på varje enskild uppgift går att koppla till delar av eller hela kriterier formulerade i ett betygssystem. Ett elevsvar värderas med en av nivåerna i betygssystemet. Fram till revisionen av läroplan och kursplaner 2011 var kriterierna formulerade för att definiera betygsskalan Ej Uppnått Målen (EUM), Godkänd (G), Väl Godkänd (VG) och Mycket Väl Godkänd (MVG). Figur 1 beskriver en del av kriterierna som elevernas svar på enskilda uppgifter var tänkta att kopplas till. Från och med 2013 års prov är beteckningarna istället E, C och A efter nivåer i den nu gällande betygsskalan. Det är inte möjligt att erhålla belägg för D och B då det inte finns några formulerade kunskapskrav för dessa nivåer.

	G	VG	MVG
Naturvetenskaplig förståelse av omvärlden Den här aspekten handlar om att beskriva och använda naturvetenskapliga begrepp, modeller och teorier.	Beskriver begrepp, modeller och teorier i kända sammanhang.	Använder begrepp, modeller och teorier i nya sammanhang	Använder begrepp, modeller och teorier för att utveckla nya frågeställningar.

Figur 1. En av tre delar i en så kallad betygsmatris (Bärman m.fl., 2010)

Ett konkret exempel på användning av beläggsmodellen för generering av provbetyg i gymnasieskolans biologi står att finna i en rapport från Umeå universitet (Grelsson m.fl., 2013).

Tabell 2. Undre gräns för provbetyg i biologi

Provbetyg	Totalt antal belägg	Varav minst
Godkänt	16	
Väl godkänt	27	7 VG eller MVG
Mycket väl godkänt	35	3 MVG

Figur 2: Konkretisering av beläggsmodellen (Grelsson m.fl., 2013, s. 7)

Enligt Figur 2 ska elevsvaren ha bedömts samla 16 belägg för provbetyg G. Beläggen kan vara på vilken nivå som helst av G, VG och MVG. För VG ska eleven ha minst 27 belägg totalt varav 7 antingen på nivån VG eller MVG. För provbetyg MVG ska eleven ha totalt 35 belägg varav minst 3 på MVG-nivå. I rapporten på sidan 6 beskrivs hur varje elevsvar kopplas till betygsnivå.

> Den modell för bedömning och resultatsammanställning som används i provet tar sin utgångspunkt i definitionen att bedömning baserar sig på de olika kunskaper och förmågor som eleven visar i en bedömningssituation. Varje förmåga som eleven visat prov på ger ett belägg. Varje belägg är knutet till ett specifikt betygskriterium och kan således även vara av olika kvalitet, d.v.s. ligga på olika betygsnivåer. Provet är konstruerat för att ge eleven möjlighet att visa belägg för så många olika aspekter av betygskriterierna som möjligt på flera kvalitativa nivåer.

I citatet framgår att de kunskaper och förmågor som en elev **visar** utgör grunden i beläggsmodellen. Det synliga uttrycket för kunskap eller förmåga är föremålet för det som utgör beläggen. Den utredning som låg till grund för nya läro- och kursplaner (Regeringskansliet, 2007) betonade starkt behovet av tydliga mål och kunskapskrav för att såväl lärare, elever och föräldrar skulle förstå systemets uppbyggnad. När kunskapskraven formulerades valdes en modell där direkt observerbara beteenden utgör objektet för lärarens bedömning och betygssättning. Dessa direkt observerbara beteenden formulerades på ett begränsat antal olika sätt. Här följer några exempel från religionskunskap:

Frank Bach

- …visar det genom att…
- …för eleven xyz resonemang…
- …genom att föra xyz underbyggda resonemang…
- …genom att beskriva/förklara…
- …genom att använda…

Liknande sätt att formulera kunskapskraven fanns i alla skolämnen.
Jämfört med tidigare kursplaner regleras inte enbart **vad** eleven ska lära sig utan också **hur** eleven ska *visa* sina kunskaper och förmågor samt vilka **kvalitetsnivåer** läraren ska finna i det eleven *visar*. Sammantaget pekar detta i en riktning mot större tydlighet men också mot begränsningar i hur elevers prestationer kan värderas. En avgörande del i ett målstyrt system är att det är möjligt att formulera målen så tydlig att det är möjligt att mäta dem och veta i vilken utsträckning de har uppnåtts (Fancourt, 2017).

Beläggsmodellen implementerades i samtliga NO- och SO-prov i grundskolans årskurser 6 och 9 från och med utprovningsomgången 2013, men med vissa skillnader mellan NO och SO. Kursplanerna med tillhörande kunskapskrav i SO-ämnena från 2011 kom att, utöver villkor om ett visst antal belägg per nivå, också att innebära villkor om belägg på viss nivå inom olika delar av kursplanen. Dylika delar har kommit att kallas förmågor. I kunskapskraven finns kopplingar till förmågorna i form av sex formuleringar. Uppgifterna i provet är, beroende på version, kopplade till fem eller sex av dessa formuleringar och därigenom till fem eller sex **delskalor** i provet.

Sedan beläggsmodellen introducerades i NO-proven 2009 har den kompletterats, i SO-ämnena, så att en viss nivå på beläggen måste nås för **flera** av förmågorna. Detta kan benämnas *breddvillkor*. De villkor som infördes i samband med NO-proven 2009 har erhållit termen *höjdvillkor*. Ett exempel hämtas från

bedömningsanvisningen i det frisläppta provet i religionskunskap från år 2015.[4] När det totala antalet belägg är sammanräknat och fördelningen av dessa belägg över de olika förmågorna och nivåerna är klarlagt, bestäms provbetyget utifrån principerna beskrivna i Tabell 1.

Tabell 1. Exempel på höjdvillkor (kolumnerna 2 och 3) och breddvillkor (kolumn 4) från provet år 2015

Provbetyg	Minsta antal belägg totalt	Fördelning belägg	Fördelning förmågor
E	16		Beläggen för nivå E ska vara fördelade över samtliga förmågor 1-4.
D	24	Minst 5 för nivå C, **eller** sammanlagt 5 för nivå C+A	Beläggen för nivå E ska vara fördelade över samtliga förmågor 1-4.
C	32	Minst 9 för nivå C, **eller** sammanlagt 9 för nivå C+A	Beläggen för nivå E och C ska vara fördelade över samtliga förmågor 1-4.
B	40	Minst 5 för nivå A	Beläggen för nivå E och C ska vara fördelade över samtliga förmågor 1-4.
A	48	Minst 8 för nivå A	Beläggen för nivå E och C ska vara fördelade över samtliga förmågor 1-4.

Denna förändring av beläggsmodellen till att också omfatta breddvillkor kan sägas vara en konsekvens av synsättet att elevernas prestationer skulle betygsättas på ett icke-kompensatoriskt sätt genom att alla delar av kunskapskravet sades behöva vara uppfyllda för en viss betygsnivå (för E, C eller A) eller till övervägande del för betygen D och B (Gustafsson m.fl., 2014; Skolverket, 2016). En följd av detta synsätt är att betydelsen av samband mellan olika delar i en elevs prestationer reduceras. Denna reduktion avspeglas i ovan citerad beskrivning av hur provbetygen i religionskunskapsproven genereras.

[4] https://www.gu.se/didaktik-pedagogisk-profession/nationella-prov-i-religionskunskap#Nationellt-prov-från-2015

Metoden för att generera provbetyg för proven i religionskunskap har förändrats över tid. När de **skarpa** proven introducerades 2014 genererades provbetyg genom såväl genom höjd- som breddvillkor. Efter några år försvann breddvillkoren och endast höjdvillkoren finns nu kvar. Därmed överensstämmer beläggsmodellen i religionskunskap numera med den ursprungliga som infördes 2009 i NO-ämnena, förutom att det nu rör sig om fler betygssteg.

Eftersökningar av studier med beläggsmodellen i fokus har så här långt inte resulterat i några funna rapporter eller artiklar. Däremot finns studier som tar ett mer övergripande grepp om reformförändringarna med fler nationella prov efter 2011 (Lidar m.fl., 2017; Lidar m.fl., 2020). Ett delresultat av dessa är att en del lärare har funnit beläggsmodellen mycket användbar i det dagliga arbetet med att planera, genomföra och utvärdera undervisning samt utvärdera elevernas lärande. Införandet av beläggsmodellen i nationella prov har således potential för att förändra lärares praktik, vilket, i kombination med avsaknad av studier av konsekvenser av beläggsmodellen i nationella prov, motiverar en studie av validitet i form av tolkning och konsekvenser av beläggsmodellens användning i de nationella proven (Kane, 2013, 2016; Messick, 1989a).

2. Syfte och forskningsfrågor

Det övergripande syftet med artikeln är att bidra med kunskap om vilka konsekvenser beläggsmodellen kan ha för olika elever och med alternativa modeller. Detta görs i denna artikel genom att å ena sidan studera utfallet för enskilda elever och å andra sidan pröva hur några, med utgångspunkt i olika tolkningar av betygssystemet och av kursplanen i religionskunskap, modeller passar de data som föreligger i de nationella proven i religionskunskap.

En modell utgår från betygsskalans endimensionella konstruktion. Den går från lägsta betyg F till högsta betyg A och signalerar att ju mer en elev kan i ett ämne desto högre betyg bör den få. En annan modell utgår från att eleven ska prövas mot varje del i kunskapskravet för sig, utan samband med övriga delar av kunskapskravet, på ett icke-kompensatoriskt sätt (Gustafsson m.fl., 2014). En tredje modell, i motsats till den förra, utgår från att de kunskaper som prövas mot varje del av kunskapskravet

samvarierar på ett kompensatoriskt sätt. Beläggsmodellen befinner sig någonstans mellan dessa två, i vissa fall närmare den första och i andra närmare den andra, beroende på hur hårda krav som sätts i bredd- respektive höjdvillkor.

I kursplanen för religionskunskap är det därutöver möjligt att urskilja två andra möjliga komponenter hos elevernas kunskaper; en om att använda faktakunskaper om världsreligionerna för *analyser* av olika slag och en där eleverna ska *resonera* om etik, moral och livsfrågor (Franck, 2017). En fjärde modell med två eller tre kunskapskomponenter, beroende på årgång av provet prövas. En femte och sista modell prövas där de två komponenterna *analysera* och *resonera* utgör bifaktorer till en övergripande religionskunskapsfaktor som bidrar till elevernas svar på alla uppgifter inklusive *källkritik*.

Den första forskningsfrågan tar sin utgångspunkt just i att det är angeläget att veta mer om beläggsmodellens konsekvenser för provbetygen:

1. Bidrar beläggsmodellen i religionskunskapsproven med reliabel information utöver totalsumman? Om så är fallet hur är denna information beskaffad?

Beläggsmodellen utgår från några specifika utgångspunkter redovisade i citatet ovan från Grelsson m.fl. (2013) Med denna utgångspunkt formuleras andra forskningsfrågan:

2. Hur väl passar fem olika, utgående från olika tolkningar av kursplanen för religionskunskap, modeller om elevers kunskaper i religionskunskap, resultaten från de nationella proven?

Beroende på utfallen av modelleringarna kan alternativa modeller för att generera provbetyg övervägas. En tredje forskningfråga utgår från resultaten på forskningsfråga två:

3. Hur skulle en modell för provbetyg kunna konstrueras med utgångspunkt från resultaten av modelleringarna?

Med dessa tre forskningsfrågor som vägledning tar avsnittet om teori och metod vid.

3. Teori och metod

Utgångspunkten är att en elevs kunskaper inte är direkt observerbara utan endast uppskattas indirekt, t.ex. genom resultat på prov. Storheter som inte kan observeras direkt betecknas latenta. Latenta storheter kan dock ge upphov till - orsaka - olika manifesta tecken på sin existens. I fallet med proven i religionskunskap är elevens svar på uppgifterna sådana manifesta uttryck. Dessa uttryck värdesätts med någon form av beteckning för att markera kvaliteten på svaret. De manifesta uttrycken, eller indikatorerna, kan användas för att dra slutsatser om den latenta variabeln (Beaujean, 2014, s. 37–39; Loehlin, 2004, s. 1–4; Wilson, 2005). För att kunna göra detta behövs ett resonemang, en argumentation som gör det troligt att de indikatorer som erhålls om elevens kunskaper är tecken på det just det som önskas. Enligt Messick är inte denna fundamentala aspekt av validitet tillämplig för en enskild indikator (Messick, 1989b; tolkad i Gustafsson, 2008). Istället behövs relativt många indikatorer för att kunna dra valida slutsatser om elevens kunskaper.

Det är möjligt att med hjälp av konfirmatorisk faktoranalys formulera modeller som kan prövas mot data. Brunner m.fl. (2012) presenterar fyra principiellt olika konfirmatoriska modeller, a) Enfaktormodellen som endast innehåller en generell faktor som speglar en övergripande egenskap som förklaring till resultaten, b) en första ordningens modell med flera faktorer som fokuserar på specifika egenskaper. De två sista modellerna kombinerar idén om en generell faktor med idén om flera olika faktorer, men där c) har en övergripande faktor som influerar på de underliggande och d) har en faktor på motsvarande sätt som enfaktormodellen men kombinerat med flera faktorer. Denna sistnämnda går under beteckningar som *nested-factor model* (NF), *general-specific model* och *bifactor model* (Chen m.fl., 2006; refererad i Brunner m.fl., 2012).

På senare tid har bifaktormodellen rönt allt större uppmärksamhet (Reise, 2012) och antalet artiklar där bifaktor nämns i titeln har ökat från några enstaka år 2000 till runt 180 år 2018 (Zhang m.fl., 2021). I bifaktormodellen kan den generella faktorn betraktas som en bred övergripande faktor och de avgränsade deldomänerna som specifika faktorer.

3.1 Data

Data utgörs av de provresultat som lärare har skickat in över hur varje enskilt elevsvar på varje uppgift i proven har bedömts. Eleverna väljs efter födelsedatum i månad. Antalet födelsedagar är valt för att erhålla cirka 2000 elevers resultat på provet. Tillvägagångssättet bör leda till ett slumpmässigt och representativt urval. Dylika insamlingar sker rutinmässigt av de institutioner som utvecklar de nationella proven.

Tabell 2: Urvalens storlek med fördelning mellan flickor och pojkar

År	Flickor	Pojkar	Ej angivet	Totalt
2014	863	891	333	2087
2015	994	995	944	2933
2016	912	906	718	2536
2017	744	786	895	2425
2018	947	1013	0	1960
2019	1117	1211	0	2328

Lärarna har rapporterat in så kallade *belägg* för betygen F, E, C eller A för utvalda elever och samtliga uppgifter i proven. Data föreligger alltså på ordinalskala för alla uppgifter i proven. Beläggen har omvandlats till poäng 0, 1, 2 och 3 för att underlätta beräkningar, dock fortfarande på ordinalskala. Matematiskt innebär detta ingen skillnad mot F, E, C och A. All tillgänglig data har använts i alla analyser utom i de fall där uppgifter saknas. Det sistnämnda gäller jämförelser mellan pojkar och flickor där denna information saknas för vissa årgångar för en andel av eleverna (se Tabell 2). I övrigt inga bortfall eller annan reduktion av data. Skillnaden i antal mellan åren beror med största sannolikhet mest på en variation i följsamhet att följa uppmaningen att skicka in resultat och i andra hand på att antalet födda de utvalda datumen.

3.2 Procedurer

Etablerade metoder för deskriptiv statistik har använts som underlag till tabeller och diagram (Crocker & Algina, 2008). Utöver det allra det mest frekvent förekommande används också begreppen *decision accuracy/consistensy* (eller *classification accuracy/consistensy*) (Livingston, 2018; Meyer, 2010) för att studera beläggsmodellens konsekvenser. Med dessa begrepp avses, något förenklat, i vilken utsträckning, på grund av till exempel slumpfel, eleverna erhåller för höga eller för låga provbetyg i förhållande till deras faktiska kunskaper. En metod, introducerad av Haberman (2008) används för att undersöka ifall delskalorna i provet tillför något utöver totalpoäng på provet som helhet. Metoden har utvecklats med signifikanstestning av resultaten av Sinharay (2019) och har implementerats i ett tillägg till R (Dai m.fl., 2022). För att avgöra om data är lämpliga för faktoranalys har Kaiser-Meyer-Olkins Measure of sampling Adequacy[5] (Kaiser & Rice, 1974) och Bartlett's test of sphericity[6] använts (Bartlett, 1951).

De fem modellerna illustreras med en av versionerna (årgångarna) av provet, medan resultaten redovisas för alla prövade årgångar. Denna del av resultatredovisningen avslutas med en granskning av hur en av modellerna fungerar för undergrupperna flickor och pojkar tillsammans med hur väl modellen förutsäger resultaten för dessa grupper. Detta görs med hjälp av en procedur med allt striktare mätvillkor för skillnader mellan grupperna (Brown, 2015, s. 241–267).

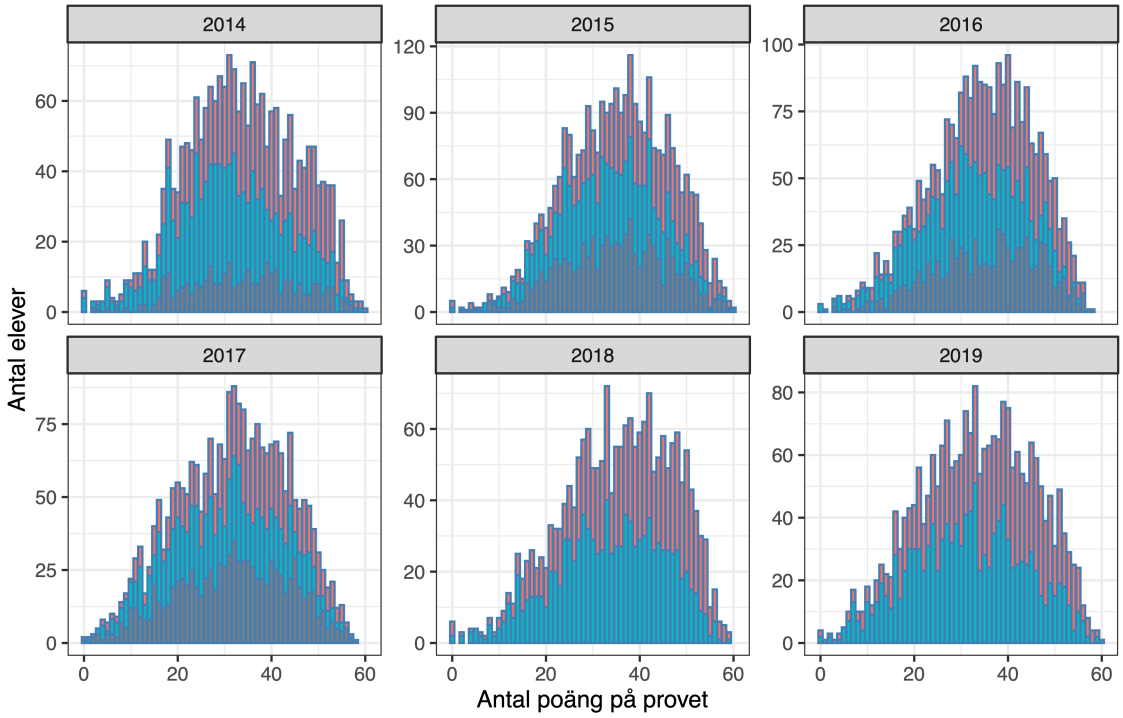
Inledningsvis sammanställs resultaten i form av sex diagram där totalt antal elever med en viss totalpoäng på proven visas med den sammanlagda höjden på varje enskild stapel.

---

[5] Kaiser-Meyer-Olkins Measure of sampling Adequacy (KMO) används för att bedöma hur lämpligt det är att använda faktoranalys för den tänkta datamängden. KMO-värdena ligger mellan 0 till 1, där värden närmare 1 anger större lämplighet för faktoranalys. Värden större än 0,8 anses vara bra.

[6] Bartlett's test of sphericity jämför en observerad korrelationsmatris med enhetsmatrisen (enbart ettor i diagonalen). Testet går ut på att se om det finns en avvikelse mellan enhetsmatrisen och korrelationsmatrisen för aktuell data. Om skillnaden är signifikant ($p<0,05$) är villkoren för faktoranalys uppfyllda.

Rödaktig färg visar resultat för dem som har angetts som flickor, turkos färg för pojkar och den mörkare färgen står för dem där biologiskt kön inte har angetts.



Figur 3: Poängfördelning för Religionskunskapsproven 2014-2019 (gruppvis kumulativt)

Gemensamt för alla sex versionerna är att pojkars resultat utgör en större andel av de lägre poängsummorna i diagrammen och tvärtom för flickorna. Resultatfördelningarna liknar varandra genom att ha drag av en normalfördelning men med en negativ skevhet (svans åt vänster). För resultat från proven som helhet behandlas data som om de vore på kvotskala, även om ursprunget till summorna är på ordinalskala. Antal insamlade resultat, medelvärde, median, standardavvikelse och skevhet framgår av Tabell 3.

Tabell 3: Sammanfattande beskrivning av de nationella proven i religionskunskap 2014-2019.

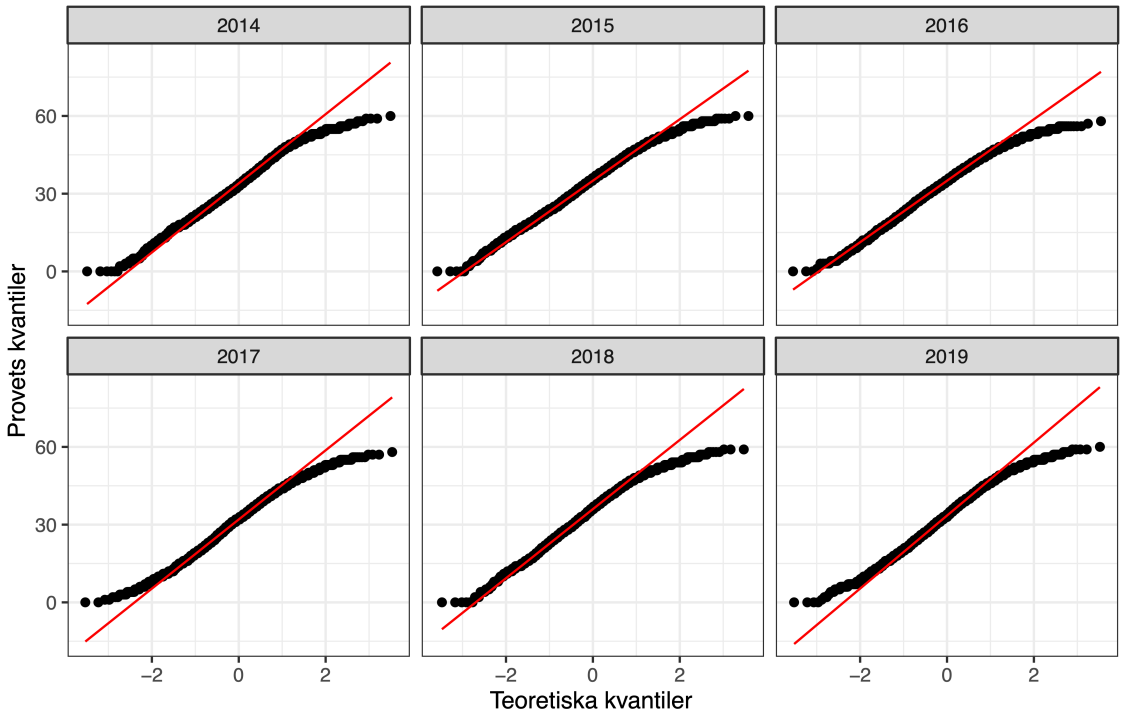
År	Antal	Medel	Median	Min	Max	SD	Skevhet*	Alpha
2014	2087	33,44	33	0	60	11,76	-0,13	0,92
2015	2933	35,16	36	0	60	11,06	-0,22	0,92
2016	2536	34,30	35	0	58	10,98	-0,37	0,92
2017	2425	31,64	32	0	58	11,79	-0,20	0,93
2018	1960	35,33	36	0	59	11,73	-0,37	0,93
2019	2328	33,29	34	0	60	12,13	-0,19	0,93

*Justerad Fisher-Pearson koefficient för skevhet

Antalet elevresultat som har samlats in varierar mellan 1960 år 2018 och 2933 år 2015 med lägst resultatmedelvärde 2017 och högst 2018. Antalet möjliga poäng på proven varierar något över åren men ligger runt 60. I Tabell 3 anges inte detta utan istället den högsta poäng någon elev har erhållit varje år. Reliabiliteten hos provet ligger en bit över 0,9 (Chronbachs alfa[7]) för alla sex årgångarna. Värdena för Skevhet visar att poängfördelningarna för alla sex proven att det finns en svans åt vänster eller annorlunda uttryckt att medelvärdet ligger något högre än mitten (se Figur 3).

För att avgöra om poängfördelningarna är normalfördelade jämförs de med en normalfördelning genom att rita QQ-diagram. Detta görs i Figur 4.

[7] Ett mått på reliabilitet som ligger mellan 0 och 1, ju högre desto högre reliabilitet. Värden över 0,9 anses som nödvändiga vid mätningar som har avgörande betydelse för individer.



Figur 4: Kontroll av normalfördelning av proven 2014 - 2019

Ett QQ-diagram jämför kvantiler från en fördelning (t.ex. ett prov) med kvantiler[8] från en annan, till exempel en normalfördelning. Om dessa bildar en rät linje är de två jämförda fördelningarna från jämförbara populationer. Här har de nationella proven i religionskunskap jämförts med en normalfördelning och följer approximativt en sådan, möjligen förutom nära minimum och maximum.

Kaiser-Meyer-Olkins Measure of sampling Adequacy har ett värde runt 0,97 för samtliga sex versioner av provet. Bartlett's test of sphericity visade ett signifikant resultat ($p<0{,}001$) för alla sex versioner av provet. Sammantaget visar detta att data är lämpliga för faktoranalys.

[8] En kvantil delar en fördelning i regelbundna intervaller. Vanliga kvantiler är kvartiler och procentiler.

Samtliga modeller värderas mot följande ungefärliga tumregler: RMSEA[9] helst mindre än 0,06; TLI[10] helst större än 0,95; CFI[11] helst större än 0,95; SRMR[12] helst mindre än 0,08, allt tillsammans med kvalitativa överväganden (Hu & Bentler, 1999). All analys av data och textproduktion har genomförts i RStudio (RStudio Team, 2020) med R (R Core Team, 2020) och tilläggen lavaan (Rosseel, 2012), psych (Revelle, 2020), equaltestMI (Jiang & Mai, 2021) och semTools (Jorgensen m.fl., 2020). Tabeller har producerats med tidyverse (Wickham m.fl., 2019) och broom (Robinson m.fl., 2020). Diagram och figurer har skapats med ggplot2 i tidyverse och semPlot (Epskamp, 2019).

4. Resultat

4.1 Fördelning av provbetyg som en följd av beläggsmodellen

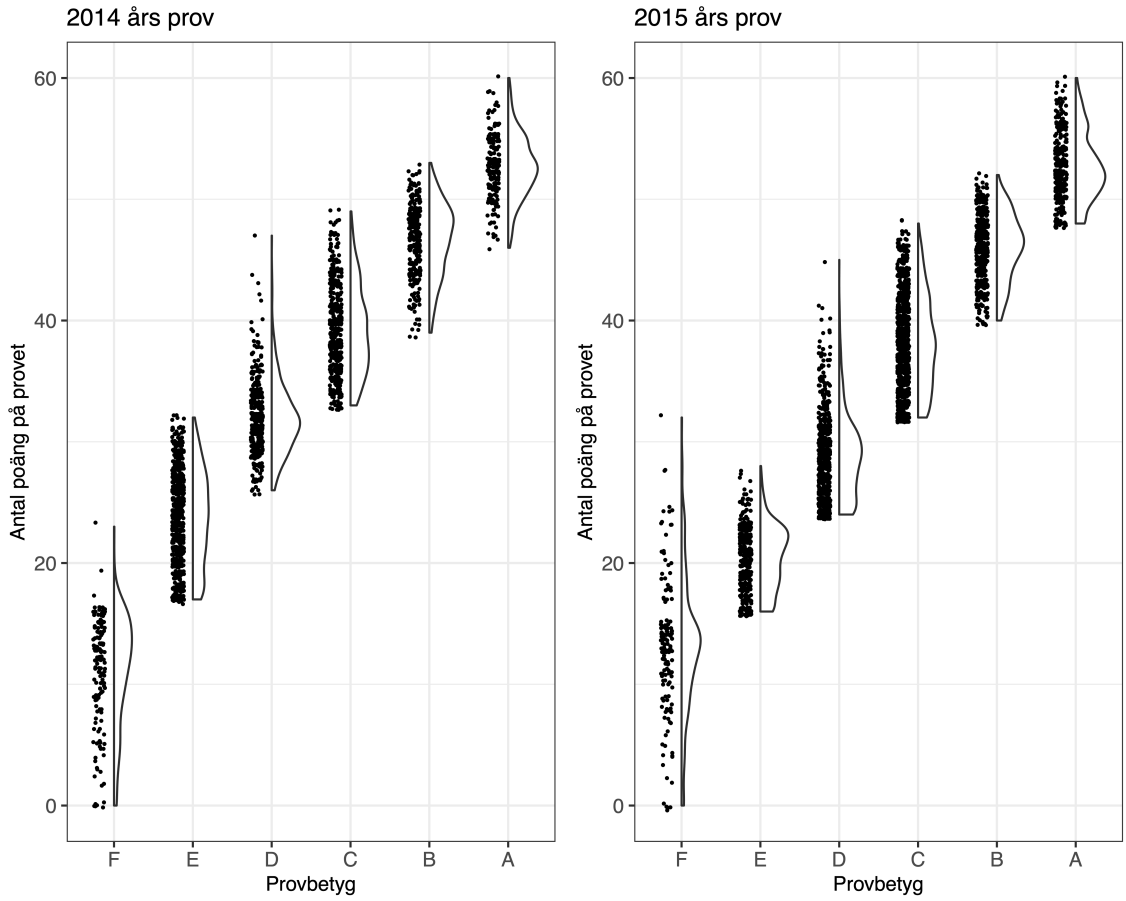
Utfallet av beläggsmodellen för de första två åren illustreras av diagrammen i Figur 5 som består av kombinationer av punkt- och violindiagram, ett för år 2014 och ett för år 2015. Violin-diagrammen illustrerar densiteten poäng inom grupp med tjockleken och spännvidden över poängskalan genom längden, medan punktdiagrammen visar antalet.

[9] Root Mean Square Error of Approximation är ett mått på skillnad mellan den observerade och den modellskapade kovariansmatrisen per frihetsgrad.

[10] Tucker-Lewis index räknas ut genom att ta RMSEA för modellen och dividera det med RMSEA för nollmodellen, vilket oftast resulterar i ett värde mellan 0 och 1, där högre värden är ett tecken på bättre modellapassning.

[11] Comparative Fit Index (CFI) används också för att jämföra hur bra en modell passar en modell utan relationer mellan variablerna, men genom att jämföra modellens med nollmodellens negativa log-sannolikhet. CFI ligger mellan 0 och 1, där högre värden indikerar en bättre anpassning.

[12] Standardized Root Mean Square Residual (SRMR) är ett index för genomsnittet av standardiserade residualer mellan den observerade och den hypoteserade kovariansmatrisen.

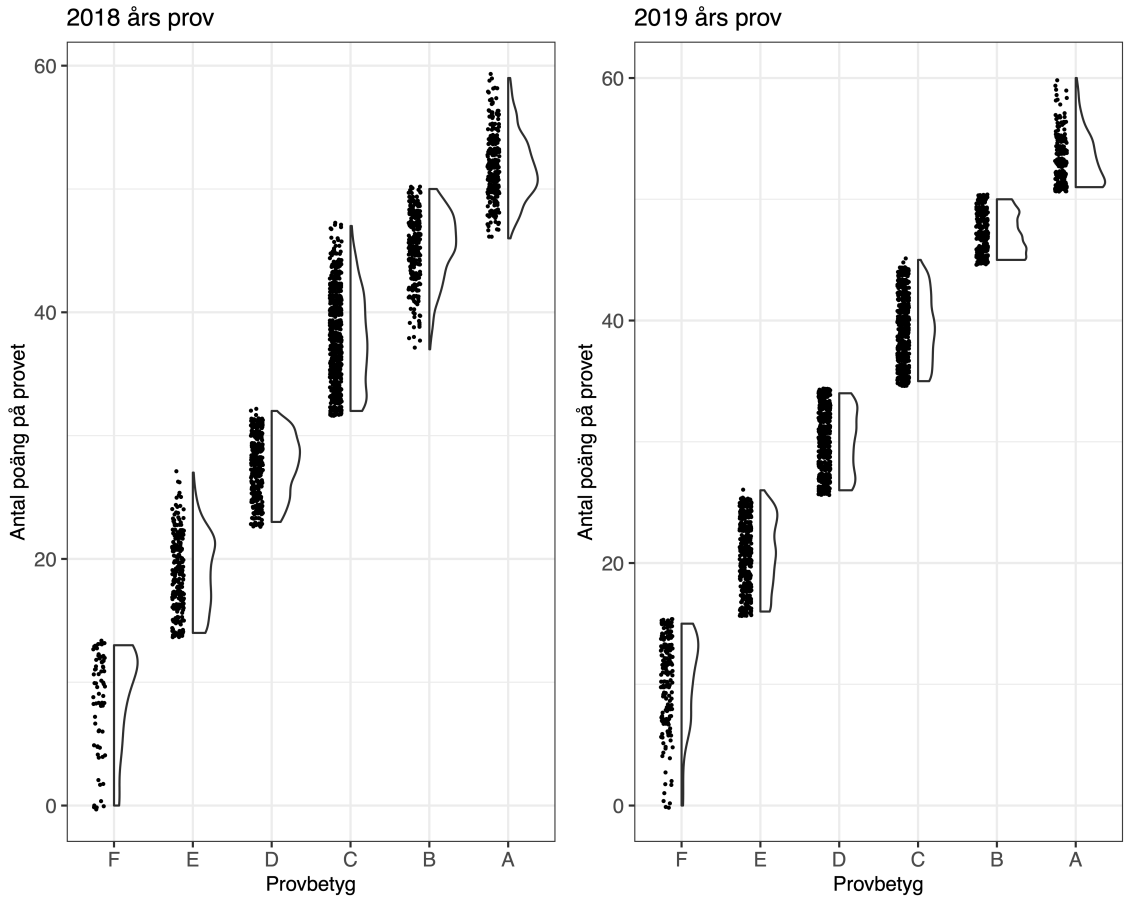


Figur 5: Fördelning av provpoäng över provbetyg för proven 2014 och 2015

I båda proven överlappar provbetygen antalet poäng på provet. Det innebär att två elever kan ha erhållit samma antal poäng (belägg) på provet, men olika provbetyg. I dataurvalet från 2014 års prov kan två elever med samma poäng på provet ha erhållit allt från D till A i provbetyg. Det rör sig dock om enstaka elever. I nationen som helhet finns det dock högst sannolikt fler. Vanligare är att elever med samma antal poäng på provet har erhållit ett eller två stegs skillnad i provbetyg. Från provet år 2015 finns det elever som har samma poäng på provet som helhet, men där provbetygen går från D till B. I samma årgång finns det också elever med samma totalpoäng på provet som har erhållit provbetyg från F till C. I urvalet finns det enstaka elever som har erhållet mer än hälften av antalet möjliga poäng (belägg) på provet, men som ändå har erhållit provbetyg F. Vid gränsen mellan F och E noteras att villkoret att eleven måste erhålla E-nivå för flera delkunskapskrav innebär att elever med

poängantal klart över lägsta möjliga för provbetyg E har erhållit provbetyg F. Det är fler elever som påverkas av breddvillkoret för gränsen mellan F och E år 2015 än 2014. Därutöver är det också ett större överlapp år 2015 än år 2014.

I Figur 6 förekommer inga överlapp i antalet poäng som eleverna har erhållit vid gränsen mellan E och F. Detta beror på att de så kallade breddvillkoren, vid den här tiden, har tagits bort. Däremot fanns höjdvillkoren kvar vilket noteras genom att poängsummorna överlappar de högre provbetygen, åtminstone i provet från 2018. Där har samma poäng på provet för en viss andel av eleverna genererat allt från provbetyg C till A.



Figur 6: Fördelning av provpoäng över provbetyg för proven 2018 och 2019

Provbetygsgränserna för provet från 2019 har satts så att höjdvillkoren har fått mycket liten betydelse. Poängsummorna på provet överlappar nästan ingenting mellan de olika provbetygen. Förändringarna av procedurerna vid kravgränssättning över tid speglar sannolikt ett ökat behov av reliabilitet i mätningen och av att minska variationen i andelarna provbetyg från år till år.

4.2 Delskalorna

För att veta om de olika delskalorna i provet bidrar med något utöver totalskalan användes Habermans metod (Haberman, 2008; Sinharay, 2019). Beräkningarna visar att en delskala har tillfört värdefull information i två versioner av provet, nämligen årgångarna 2015 och 2017. I båda fallen rör det sig om delskalan som knyter an till följande formulering i kunskapskraven: "Dessutom för eleven * resonemang om likheter och skillnader inom och mellan några religioner och andra livsåskådningar", där * står för olika kvalitetstuttryck kopplade till betygsskalan. I de sex versionerna av provet i religionskunskap finns sammanlagt 32 delskalor varav två skulle kunna ge information utöver totalpoängen i proven.

4.3 Konfirmatorisk faktoranalys

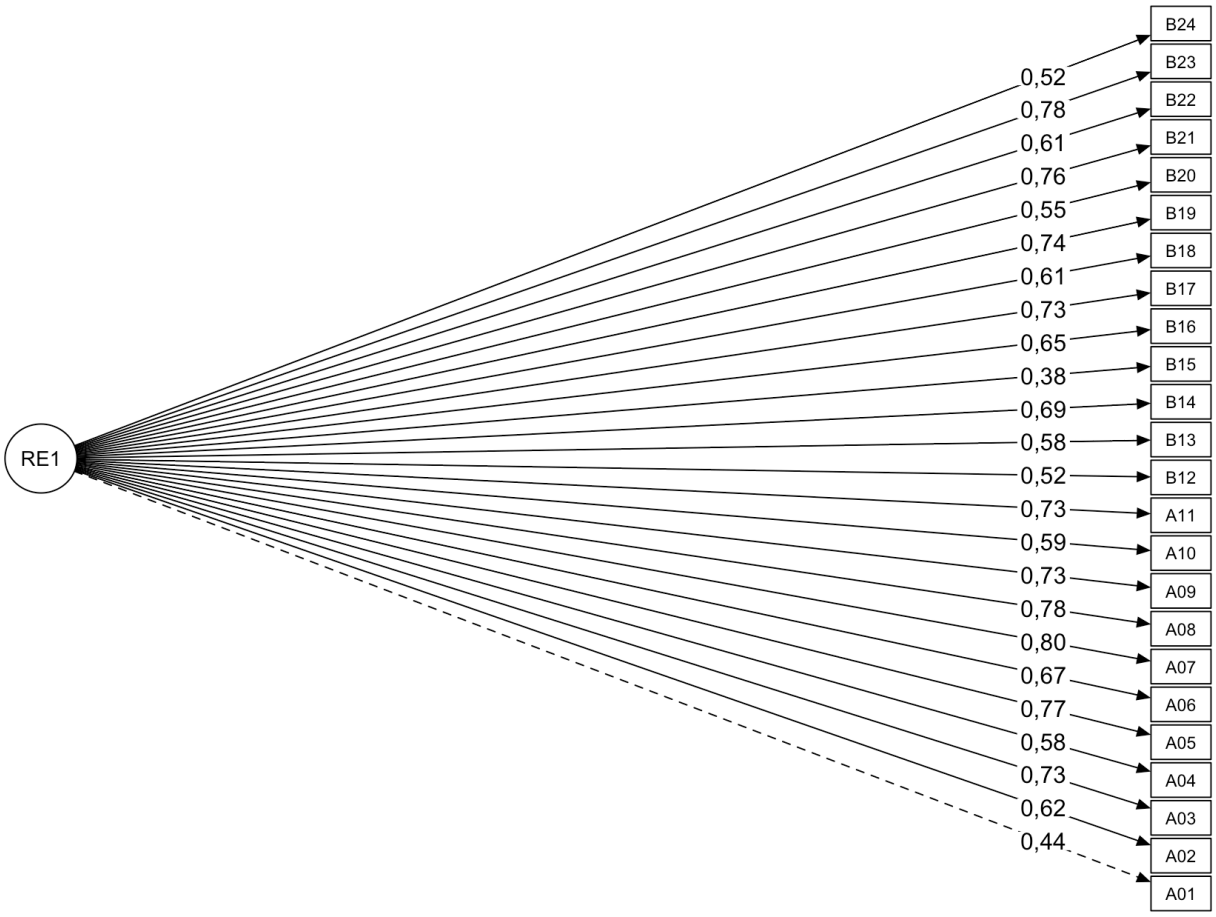
4.3.1 Modell med en latent variabel som ger upphov till svaren på uppgifterna

Den enklast tänkbara modellen har endast en latent variabel. Denna kan benämnas som "Elevens kunskaper i religionskunskap" och denna antas generera elevens svar på uppgifterna i provet. Det är rimligt att kontrollera om en sådan enkel modell är god nog. Hur väl data stämmer med denna modell för de olika versionerna av det nationella provet i religionskunskap anges i Tabell 4.

Tabell 4: Anpassning till data för religionsprovets versioner för modell ett

År	CFI	TLI	RMSEA	SRMR
2014	0,99	0,99	0,04	0,06
2015	0,98	0,98	0,06	0,06
2016	0,99	0,99	0,05	0,06
2017	0,98	0,98	0,07	0,07
2018	0,99	0,99	0,06	0,06
2019	0,99	0,98	0,06	0,06

Modellen passar data relativt bra för alla sex årgångarna av proven. Den latenta variabeln *Elevens kunskaper i religionskunskap* (RE1) är rimlig för att beskriva de sex versionerna av det nationella provet i religionskunskap. Den kunskap som krävs för att besvara uppgifterna i de olika versionerna av det nationella provet i religionskunskap kan i huvudsak beskrivas som endimensionell. Modellen tillämpad på data för sex nationella prov illustreras med 2016 års version i Figur 7.



Figur 7: Modell med en latent variabel för provet 2016

Den latenta variabeln *Elevens kunskap i religionskunskap* inverkar relativt högt på de flesta av de manifesta variablerna (bedömningen av elevsvaren på uppgifterna i proven, A01-B24). Korrelationerna mellan den latenta variabeln och de observerade prestationerna (manifesta variabler) varierar mellan 0,38 (B15) och 0,8 (A07). Värdena för RMSEA och SRMR antyder att modellen som beskriver elevernas kunskaper är god nog, men bör kunna förbättras något med andra modeller.

4.3.2 Modell utgående från att elevens svar som rör ett delkunskapskrav är oberoende av svar på uppgifter från andra delkunskapskrav

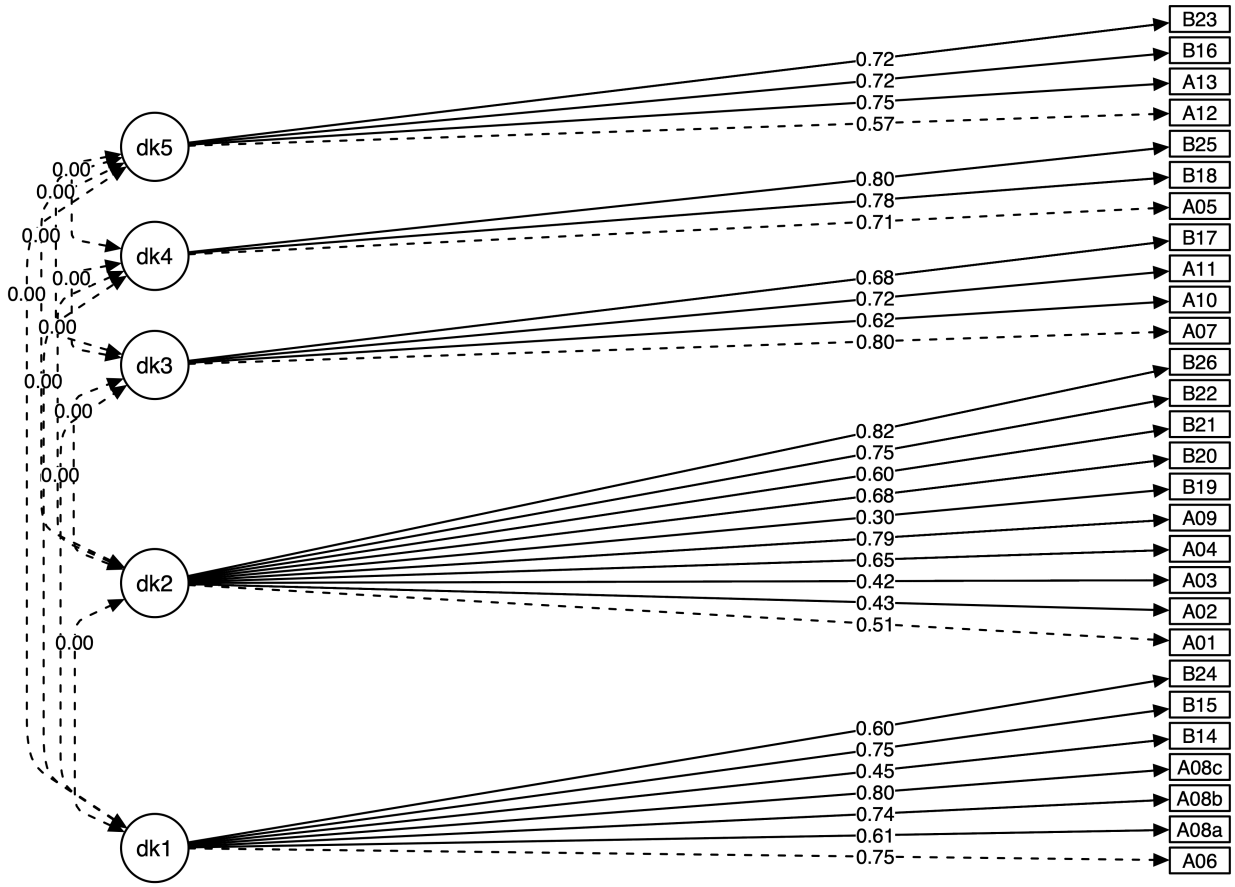
Denna modell utgår från en tolkning av skolförordningen som innebär att betygssättning av elevens kunskaper ska vägas mot varje del av kunskapskravet för sig. Uppgifterna i det nationella provet i religionskunskap har delats in efter de sex delar som kunskapskravet består av. För exempelvis provet 2019 innebär det fem uppgifter hör samman med första meningen i kunskapskravet, sex med andra, tre med tredje, fyra med fjärde, tre med femte och slutligen två med sjätte och sista delen av kunskapskravet. För var och en av de tre beskrivna betygsnivåerna; E, C och A består varje del av kunskapskravet av en mening, totalt alltså sex meningar för varje betygsnivå.

Om en elevs provbetyg ska avgöras för varje delkunskapskrav för sig innebär detta att elevens svar från någon del av provet inte ska kompensera för elevens svar i någon annan del av provet. I en konfirmatorisk faktoranalys skulle detta kunna motsvaras av att elevsvaren på uppgifter som har kategoriserats att tillhöra en viss mening (delkunskapskrav) härrör från en latent variabel som inte tillåts samvariera med någon annan latent variabel. De är oberoende av varandra. När det gäller det nationella provet i religionskunskap de två åren 2018 och 2019 finns det två indikatorer kopplade till det sjätte delkunskapskravet. Med endast två indikatorer kommer det att finnas för lite information för att kunna beräkna alla parametrar i modellen. Modellen kommer att vara underidentifierad (Beaujean, 2014) och det är därför inte möjligt att beräkna modellen utan att införa en begränsning där laddningar för de två manifesta variablerna i denna faktor sätts lika istället för att tillåtas beräknas fritt av programvaran.

Tabell 5: Anpassning till data för religionsprovets versioner för modell två

År	CFI	TLI	RMSEA	SRMR
2014	0,21	0,15	0,38	0,36
2015	0,22	0,14	0,38	0,34
2016	0,21	0,13	0,41	0,37
2017	0,25	0,18	0,44	0,39
2018	0,21	0,13	0,48	0,41
2019	0,19	0,12	0,46	0,39

Modellen passar data mycket dåligt, vilket indikeras genom de låga värdena för CFI och TLI samt de höga värdena för RMSEA och SRMR (se Tabell 5). Sannolikt beror detta i första hand på att modellen inte tillåter de latenta variablerna som förutsäger elevernas svar knutna till delkunskapskraven att samvariera. Modellen illustreras med 2014 år version av provet som exempel i Figur 8.



Figur 8: En modell med sex av varandra oberoende latenta variabler

Nollorna till vänster i Figur 8 betyder att modellen inte tillåter några korrelationer mellan de fem olika latenta egenskaperna som eleverna **antas** besitta och som gör det möjligt för dem att besvara uppgifterna. Den dåliga modellanpassningen (som beror på villkoren i modellen med förutsättningen om icke-kompensatorisk bestämning av provbetyg) tyder på att de fem (eller sex, beroende på version av provet) latenta egenskaperna som antas ge elevernas svar i själva verket samvarierar. För att undersöka detta har poängsummor för varje delskala beräknats, nu med versionen från 2019 som exempel. I Tabell 6 redovisas korrelationerna mellan dessa delskalor, benämnda dk1-dk6 i tabellen.

Tabell 6: Korrelation mellan delskalor för provet version 2019

Delkunskapskrav	dk1	dk2	dk3	dk4	dk5	dk6
dk1	1,00					
dk2	0,77	1,00				
dk3	0,62	0,69	1,00			
dk4	0,60	0,63	0,65	1,00		
dk5	0,59	0,65	0,69	0,72	1,00	
dk6	0,58	0,63	0,66	0,63	0,67	1,00

Korrelationerna mellan delskalorna är relativt höga. Samtidigt är det viktigt att hålla i minnet att så korta delskalor, med några få uppgifter, vanligtvis har lägre reliabilitet än den långa skalan för hela provet. Den lägre reliabiliteten i delskalorna försvagar korrelationen mellan dem. I konfirmatorisk faktoranalys är utgångspunkten att de latenta variablerna orsakar svaren på uppgifterna vilket innebär att motsvarande försvagning inte återfinns i korrelationer mellan de latenta variablerna.

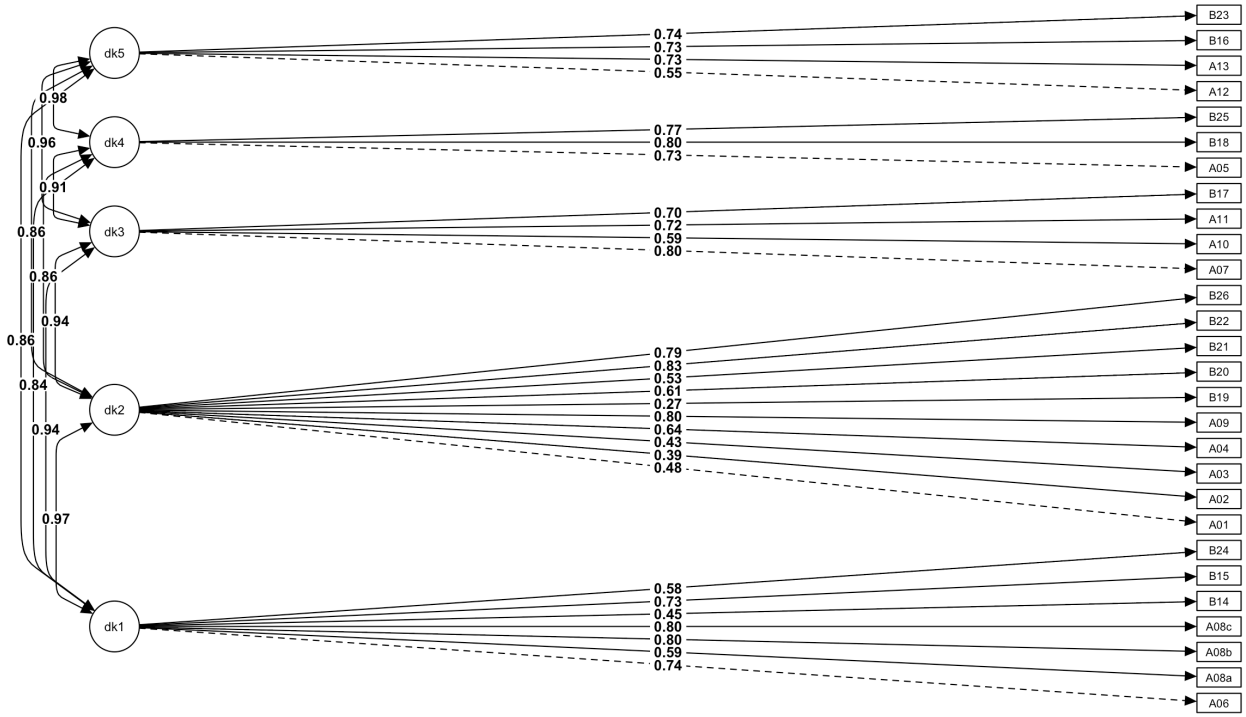
4.3.3 Modell som utgår från att elevernas svar samvarierar mellan delkunskapskraven

Det enda som skiljer denna tredje modell från den förra är att restriktionerna har lättats så att programvaran tillåts att uppskatta korrelationer mellan de latenta variablerna. Då uppvisar modellen betydligt bättre överensstämmelse med data. Såväl CFI, TLI, RMSEA som SRMR har värden som ligger på rätt sida om kriterierna för god modellanpassning.

Tabell 7: Anpassning till data för religionsprovets versioner för modell tre

År	CFI	TLI	RMSEA	SRMR
2014	0,994	0,993	0,033	0,050
2015	0,996	0,995	0,028	0,035
2016	0,992	0,991	0,043	0,047
2017	0,994	0,993	0,042	0,043
2018	0,992	0,991	0,048	0,047
2019	0,991	0,989	0,051	0,048

Det är dock en relativt komplicerad modell med, beroende på version av provet, fem eller sex latenta variabler, där korrelationerna mellan de latenta variablerna är mycket höga. I Figur 9 visas modellen och samvariationerna mellan de latenta variablerna för version 2014 av provet. Korrelationerna är delvis så höga att det är frågan om det är meningsfullt att urskilja flera av delskalorna.



Figur 9: En modell med fem mellan varandra samvarierande latenta variabler i 2014 års prov

Figuren är relativt komplicerad att läsa av med många tal placerade i närheten av varandra. Korrelationerna mellan de latenta variablerna kan lättare läsas av i Tabell 8 för 2014 års version av religionskunskapsprovet.

Tabell 8: Korrelation mellan latenta variabler i version 2014

Delkunskapskrav		dk1	dk2	dk3	dk4	dk5
	dk1	1,00				
	dk2	0,97	1,00			
	dk3	0,94	0,94	1,00		
	dk4	0,84	0,86	0,91	1,00	
	dk5	0,86	0,86	0,96	0,98	1,00

Tabellen visar att det är mycket hög korrelation mellan några av de latenta variablerna. I några fall mycket nära 1. I andra fall mellan 0,8 och 0,9. Modellen är komplicerad med fem till sex latenta variabler beroende på vilket år det rör sig om och med få manifesta variabler till varje. De höga korrelationerna innebär att de latenta variablerna var för sig inte bidrar med så mycket förklaringsvärde vilket i sin tur innebär att det troligen finns enklare modeller, med färre latenta variabler, som passar data lika bra eller bättre.

4.3.4 Modell med två eller tre korrelerade latenta variabler

Korrelationerna mellan de latenta variablerna är höga i två grupper i Tabell 8. De tre första latenta variablerna korrelerar högt med varandra och de tre sista korrelerar också högt med varandra. Det är därför rimligt att formulera en modell med två latenta variabler. Dessa bör dock relatera till en hypotes om två egenskaper hos elevernas kunskaper som rimligen kan utgöra de två latenta variablerna. En genomgång av proven 2014 - 2019 ger vid handen att uppgifterna i den ena gruppen i stor utsträckning handlar om att eleven ska använda sina faktakunskaper om världsreligionerna för att besvara frågorna. Uppgifterna i den andra gruppen kräver i större utsträckning att eleven kan föra resonemang om olika frågor som rör moral, etik och livsfrågor. Detta leder till en hypotes om två relaterade latenta variabler, en som hänger samman med elevens faktakunskaper om religioner och en annan som utgår från elevens kunskaper om etik, moral och livsfrågor. De latenta variablerna ges beteckningarna *analysera*

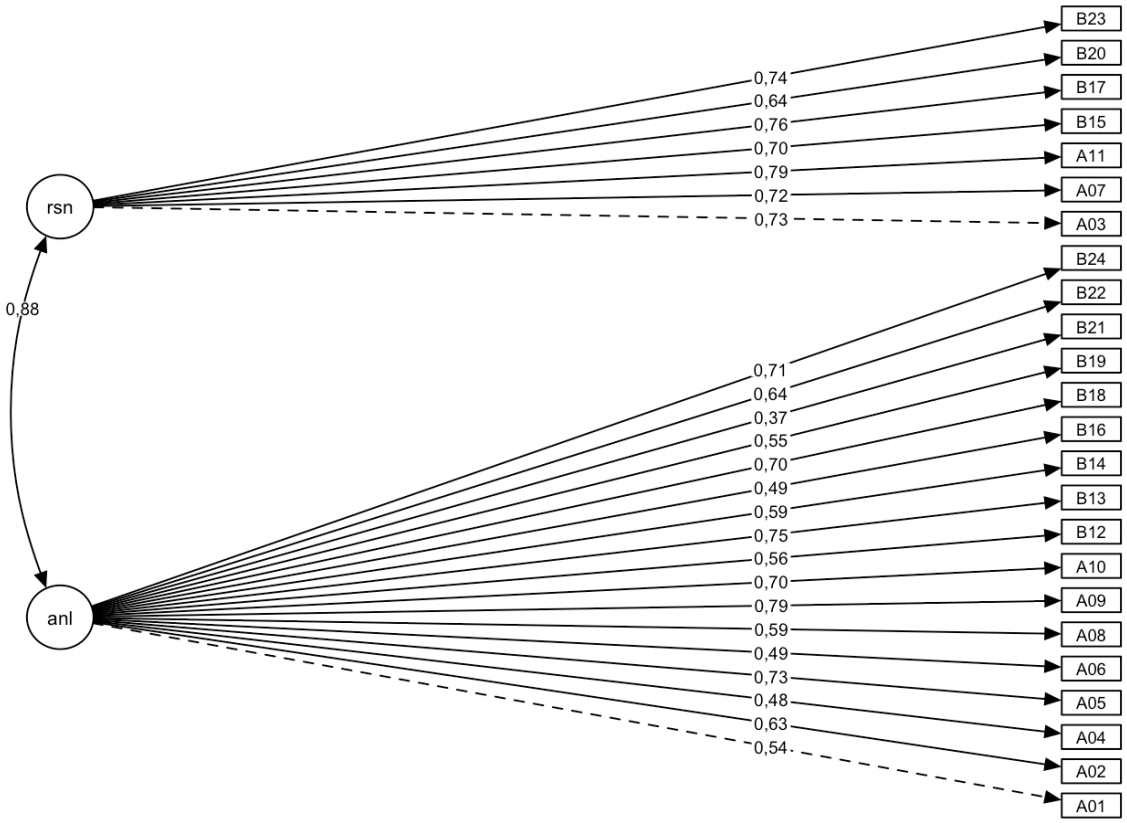
och *resonera*. Från version 2016 och framåt finns en tredje möjlig latent variabel med i proven. Den handlar om källkritik och behandlas som en tredje latent variabel för versionerna 2016 - 2019.

Denna modell, med två eller tre korrelerade latenta variabler, prövas mot data och samstämmigheten visas i Tabell 9.

Tabell 9: Anpassning till data för religionsprovets versioner för modell fyra

År	CFI	TLI	RMSEA	SRMR
2014	0,993	0,992	0,035	0,051
2015	0,985	0,984	0,052	0,055
2016	0,991	0,991	0,043	0,048
2017	0,993	0,992	0,043	0,044
2018	0,990	0,989	0,053	0,053
2019	0,988	0,987	0,057	0,055

Modellen uppvisar något bättre överensstämmelse med data än modellen med en latent variabel för samtliga valda anpassningsmått. Elevens kunskaper är i denna modell huvudsakligen uppdelade i två komponenter, där den ena relaterar till faktakunskaper och den andra till kunskaper om etik, moral och livsfrågor. En tredje komponent om källkritik finns i proven från och med 2016 till och med 2019, men det rör sig om endast två till fyra uppgifter. I Figur 10 visas att de två latenta variablerna förutsäger data från uppgifterna något bättre än modellen med en latent variabel för år 2015. Korrelationen mellan de två latenta variablerna analysera och resonera ligger på 0,88 för året 2015 och på liknande nivåer för övriga årgångar.



Figur 10: Faktorladdning med två faktorer för provet 2015

Det lilla antalet uppgifter i den latenta variabeln *källkritik* gör att mindre justeringar behöver göras i modellen för version 2018 för att modellen inte ska vara underdefinierad. Andelen genomsnittligt förklarad varians per uppgift hos de två eller tre latenta variablerna visas i Tabell 10.

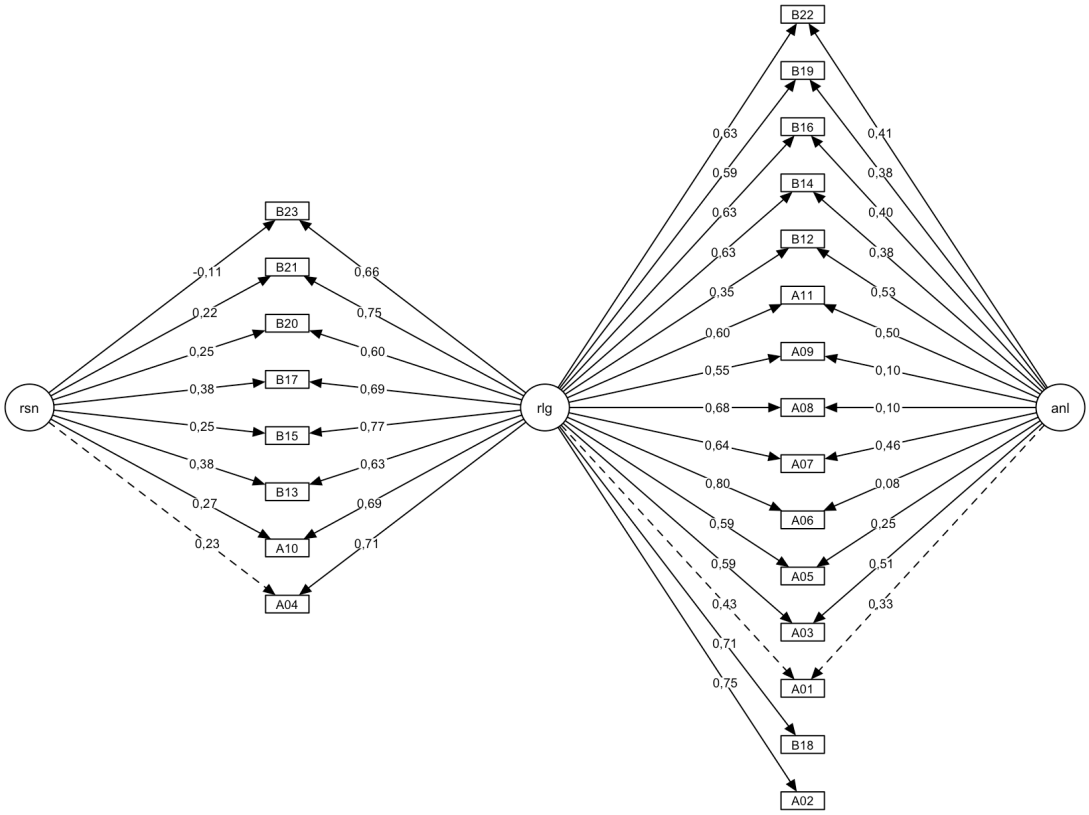
Tabell 10: Genomsnittligt förklarad varians per uppgift i de latenta variablerna

År	analysera	resonera	källkritik	total
2014	0,40	0,50	NA	0,44
2015	0,38	0,53	NA	0,42
2016	0,42	0,53	0,50	0,47
2017	0,50	0,55	0,55	0,52
2018	0,50	0,54	0,54	0,51
2019	0,47	0,51	0,52	0,48

Reliabiliteten är rimligt hög i de två delskalorna *analysera* och *resonera*, medan den är lägre i *källkritik*. Samtidigt är korrelationen relativt hög mellan de latenta variablerna. Det senare förklarar en del av den måttliga förbättringen av modellanpassning till data jämfört med en latent variabel.

4.3.5 Bifaktormodell som beskrivning av elevernas kunskaper i religionskunskap

Den sista modellen som prövas är en bifaktormodell med en huvudfaktor (latent variabel) och två bifaktorer, en för den faktaorienterade latenta variabeln *analysera* och en för variabeln *resonera*. I proven 2016 - 2019 har några uppgifter kopplade till delkunskapskravet *källkritik* tagits med. Dessa möjliggör bifaktormodellen genom att bidra med manifesta variabler så att modellen blir överidentifierad. Exempel på en grafisk representation ges med data från 2019 års prov i Figur 11.



Figur 11: Bifaktormodell av provet från år 2019

Figuren visar en generell latent variabel (rlg) som påverkar alla uppgifter och två latenta variabler (rsn och anl) som påverkar det som antas påverka utöver den generella latenta variabeln. Den ena (anl) är kopplad till *analysera* och den andra till *resonera* (rsn). De flesta laddningarna är positiva, men i det här fallet är det en uppgift i varje bifaktor med negativ laddning. Det är möjligt att förbättra anpassning till data genom att låta dessa enbart tillhöra den generella faktorn, men för att vara trogen hypotesen behålls de i bifaktorerna. I Tabell 11 redovisas hur modellen passar data för versionerna 2016 - 2019 av det nationella provet i religionskunskap.

Tabell 11: Modellanpassning för bifaktormodell för proven 2016 - 2019

År	chisq	CFI	TLI	RMSEA	SRMR
2016	374	0,999	0,998	0,017	0,031
2017	648	0,997	0,996	0,029	0,033
2018	736	0,996	0,995	0,036	0,036
2019	648	0,997	0,996	0,030	0,031

Denna modell av elevernas kunskaper i religionskunskap passar data bäst av de fem prövade hypoteserna. Genom att beräkna den genomsnittliga uppskattade variansen per uppgift går det att få en uppfattning om i vilken utsträckning bifaktorerna bidrar utöver den generella faktorn. Detta visas i Tabell 12.

Tabell 12: Genomsnittligt uppskattad varians per uppgift i de latenta variablerna för bifaktormodellen

År	Analysera	Resonera	Generell	Total
2016	0,15	0,07	0,41	0,51
2017	0,19	0,04	0,42	0,54
2018	0,12	0,07	0,45	0,54
2019	0,14	0,08	0,42	0,52

I genomsnitt tar en uppgift hand drygt 50 % av variansen i data i bifaktormodellen. Det är störst andel av de fem modellerna. Det finns också en tendens till att bifaktorn som bygger på de delar av kunskapskravet som handlar om att *analysera* med hjälp av faktakunskaper om världsreligionerna fångar mer av variansen i data än bifaktorn *resonera*.

Modellen tillämpad för undergrupperna flickor och pojkar har använts för att kontrollera hur väl modellen förutsäger resultaten för dessa grupper. Även detta har genomförts för 2019 års prov. Kolumnerna Medelpr och SDpr i Tabell 13 innehåller förutsägelser utgående från modellen och kan jämföras med de faktiska resultaten Medel och SD.

Tabell 13: Test av modellen för grupperna flickor och pojkar för 2019

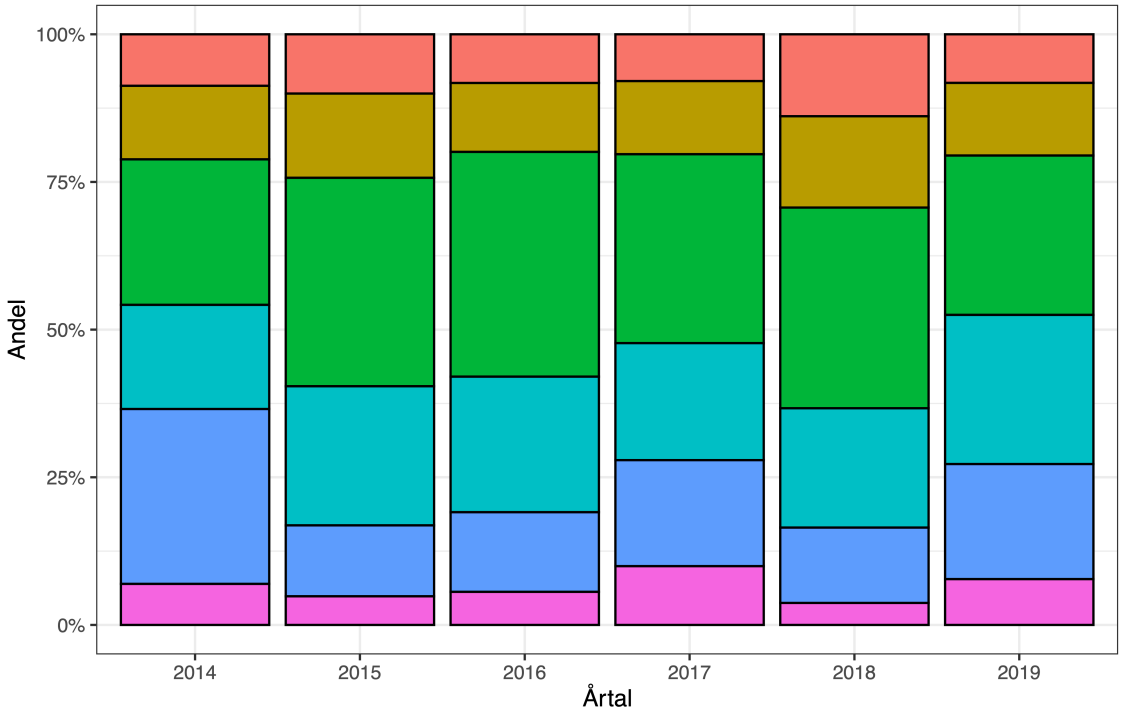
gender	Antal	Medel	SD	Medelpr	SDpr
Flickor	1117	35,37	11,90	35,38	11,16
Pojkar	1211	31,37	12,04	31,35	11,30

Skillnaden mellan modellens förutsägelse och faktiska resultat rör sig om någon hundradel när det gäller medelvärden för de två gruppernas resultat medan för standardavvikelse rör det sig om tiondelars skillnad.

Sammanfattningsvis visar de fem prövade modellerna att religionsproven i huvudsak mäter en endimensionell latent variabel, men där det också finns två små kompletterande latenta variabler. Detta resultat för vidare mot andra möjligheter att generera provbetyg än beläggsmodellen.

4.4 Generering av provbetyg

De nationella proven ska leverera provbetyg för enskilda elever varje år. Olika modeller för att generera dessa har utvecklats i proven för de olika skolämnena. På grund av hur det svenska betygssystemet är uppbyggt och tolkas ska dessa modeller fastställas och finjusteras innan proven ges. När det gäller de nationella proven i religionskunskap visas utfallet i Figur 12.



Figur 12: Provbetygsfördelning 2014-2019

Medelvärdet på proven samvarierar med provbetygen (se Tabell 3, Figur 3 och Figur 12). De år medelvärdet är högt är också provbetygen höga och tvärtom. Om detta beror av huruvida proven varierar i svårighetsgrad eller om de olika årgångarna med elever är olika duktiga är inte möjligt att säkert uttala sig om utan att etablera länkar mellan proven, antingen i form av gemensamma uppgifter mellan eller parallellt med proven. En inte alltför avancerad gissning är att svårighetsgraden mellan de olika årgångarna av proven varierar mer än kunskapsnivån hos olika årgångar av elevgrupper. Figur 12 ger en inblick i hur andelar av de olika provbetygen har varierat över de sex år proven har givits.

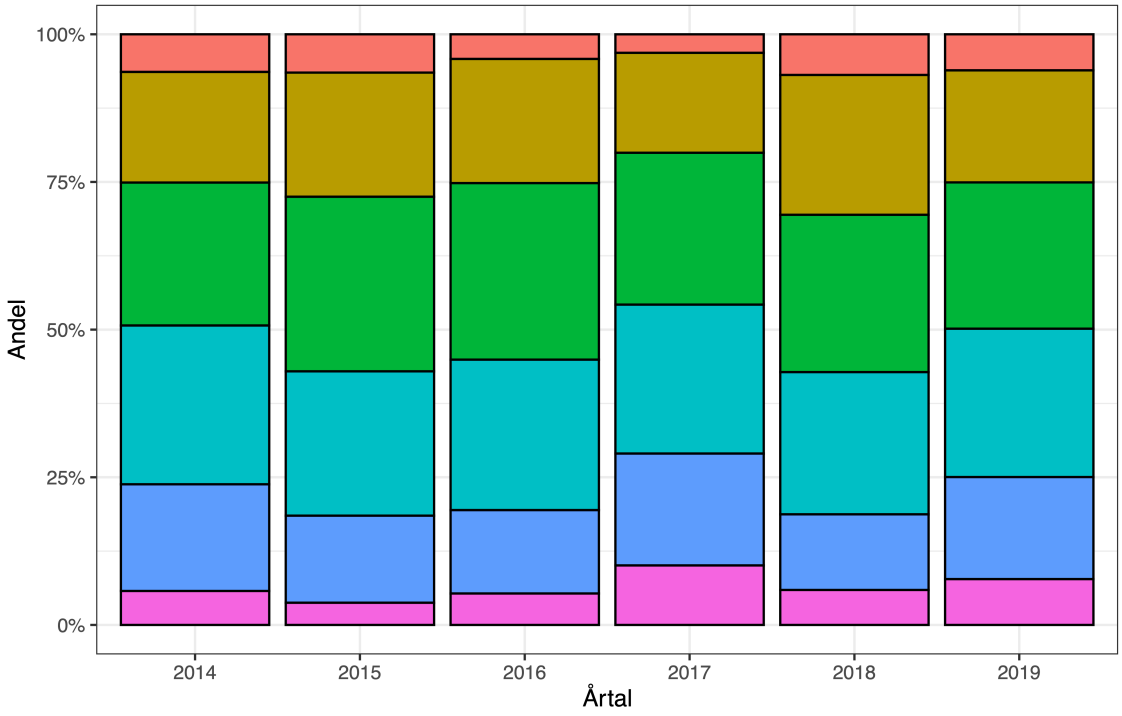
För att skapa en bild av stabiliteten hos provbetygen som härrör från beläggsmodellen i de nationella proven jämförs dessa mot utfallet av att sätta exakt samma provbetygsgränser för alla proven 2014 - 2019. Enligt Figur 3 liknar poängfördelningarna i de olika årgångarna av proven normalfördelningen. Detta gör att det är rimligt att pröva kravgränser där poängintervallen är så lika varandra som möjligt till storlek.

Då proven omfattar i storleksordningen 60 poäng och det finns sex provbetyg samt att poängfördelningen i princip är normalfördelad vore det naturligt med intervall som omfattar ungefär 10 poäng. Dock är fördelningarna lite skeva med en förskjutning ovanför mitten på skalan. Detta gör att gränserna behöver förskjutas år samma håll. Ett förslag visas i Tabell 14.

Tabell 14: Förslag på gränser för olika provbetyg

Provbetyg	**Poängintervall**
F	0-16
E	17-24
D	25-33
C	34-42
B	43-51
A	>=52

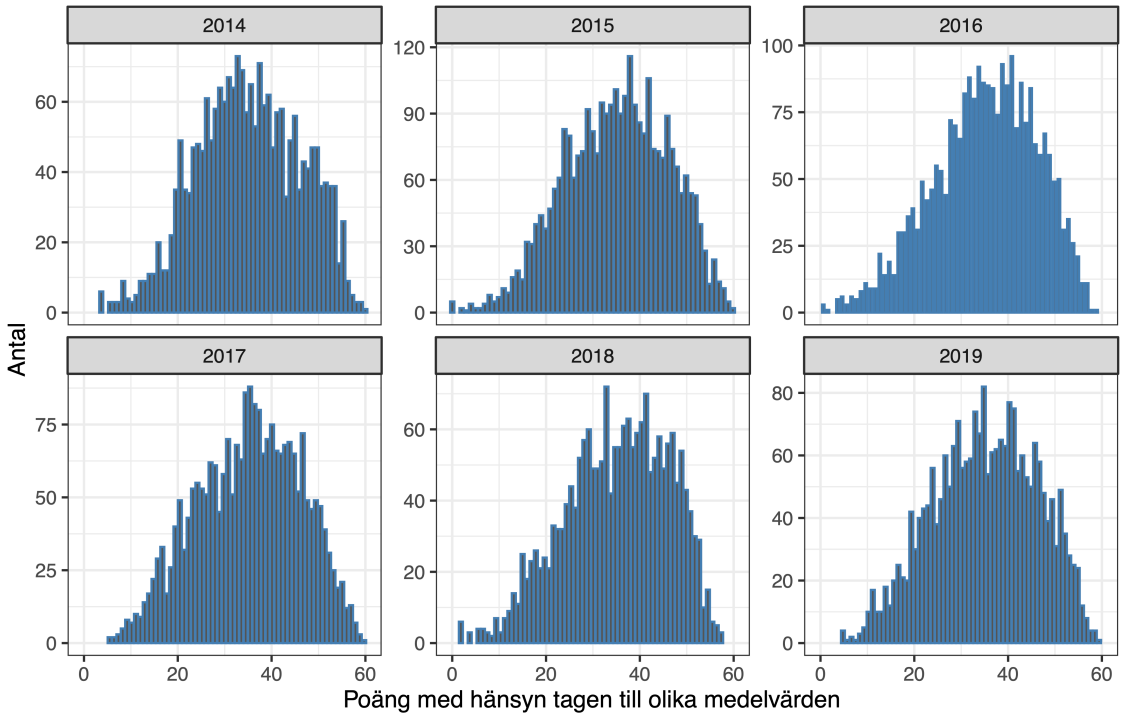
Utfallet av denna enkla princip ger en något mindre variation i provbetygen än när en separat kravgränssättning med beläggsmodellen har genomförts varje år.



Figur 13: Provbetygsfördelning med ekvidistans mellan provbetygsgränserna

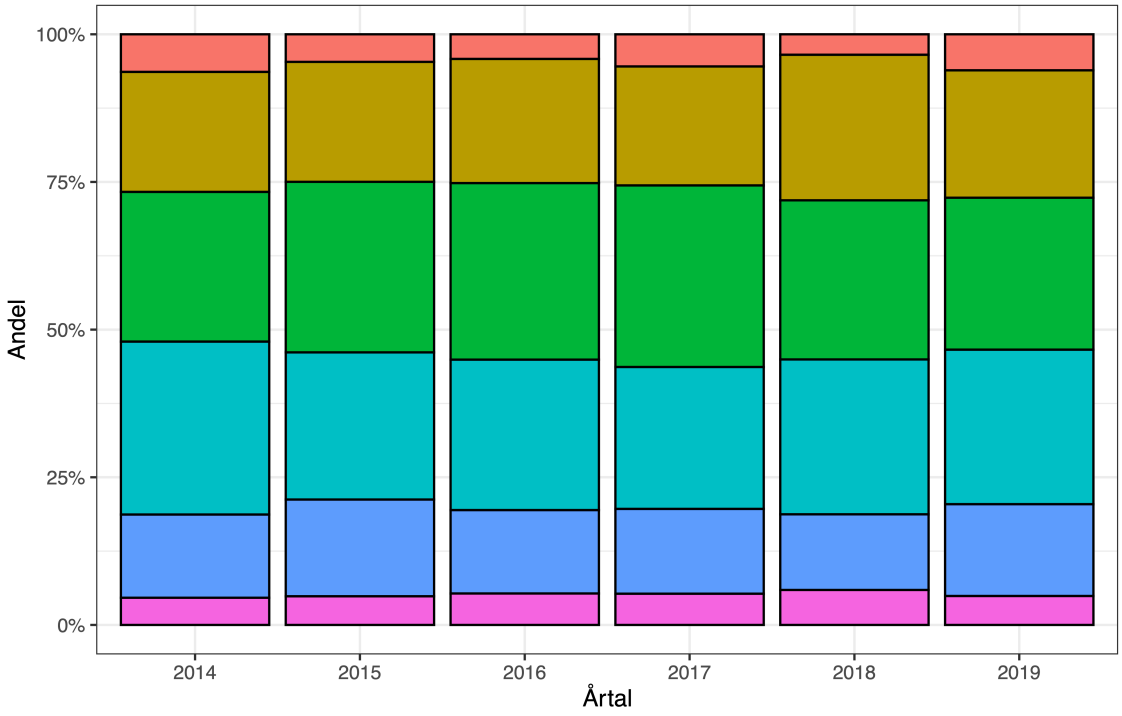
Därutöver är det rimligt att ta hänsyn till provens medelvärde. Dock är detta först möjligt efter att proven har genomförts och då skala om resultaten till en gemensam skala från 0-60 poäng. Fördelningarna förändras inte på annat sätt än att alla värden förflyttas på skalan som en slags anpassning till svårighetsgraden på proven. Provet från år 2018 visar tecken på att vara lättare än de andra. Detta syns till exempel genom en viss ansamling av resultat högt på skalan. Ett annat tecken är att det inte finns några resultat ända upp till 60 poäng. Det handlar sannolikt om en takeffekt där de bästa eleverna inte riktigt gavs möjlighet att prestera efter sina kunskaper.

Resultatfördelningarna ser alltså likadana ut som i Figur 3, men beroende på medelvärdena är de flyttade antingen åt höger eller vänster. Beroende på standardavvikelsen i proven är de också något utsträckta eller ihoptryckta utefter skalan. På det sättet blir proven mer jämförbara med varandra.



Figur 14: Poängfördelning med hänsyn tagen till olika medelvärden de olika provåren

Genom denna transformation är det möjligt att beräkna provbetygsfördelningar för proven på den gemensamma skalan.



Figur 15: Provbetygsfördelning efter att resultatfördelningarna har placerats på samma skala.

Genom dessa förhållandevis enkla åtgärder i kombination med att de sex proven håller hög kvalitet i sin helhet och att fördelningarna över poängskalan liknar varandra från år till år erhålls relativt hög likvärdighet i provbetygsfördelningarna mellan åren.

4.5 Sammanfattning av resultat

Elevernas totalpoäng från proven i Religionskunskap är normalfördelade förutom helt nära min- och maxpoäng. De uppvisar en liten skevhet där medelvärdena för proven ligger något högre än mitten på skalan. Samtliga sex prov har en reliabilitet om minst 0,92. Implementeringen av beläggsmodellen har lett till en överlapp av poäng för olika provbetyg. Det finns exempel där två elever med samma totalpoäng på provet har erhållit D respektive A som provbetyg eller F respektive C. Det finns ett exempel där eleven har 32 poäng (belägg) av 60 möjliga, men trots detta har erhållit provbetyg F. Samtidigt finns det många elever som har 16 poäng på samma prov som har erhållit E i provbetyg. Därutöver finns det flera exempel där provbetygen skiljer ett eller två steg för elever med samma

totalpoäng på provet. Konfirmatorisk faktoranalys visar att de hittills sex genomföra nationella proven i huvudsak är endimensionella kompletterade med två möjliga bifaktorer (Brown, 2015, s. 301) där den ena utgår från elevernas kunskaper för att föra resonemang om etik, moral och livsfrågor och den andra från elevernas faktakunskaper om religioner. Modellen förutsäger resultaten mycket bra. Tillräckligt hög reliabilitet, endimensionalitet och normalfördelning möjliggör ett enkelt och robust system för att sätta gränser för olika provbetyg, nämligen lika många poäng mellan de olika provbetygen. Om detta genomförs och det tillåts att ta hänsyn till varierande svårighetsgrad mellan olika versioner av proven går det att erhålla god stabilitet hos provbetygen. Till detta bör fogas att den metod som Haberman (2008) föreslår leder till att just en delskala som berör *analysera* med hjälp av faktakunskaper bidrar med information utöver poäng-resultatet från hela provet. Det ger lite extra stöd åt att denna faktor är rimlig. Dock är det endast i två versioner av sex undersökta som det framträder tydligt. Utöver detta visar Tykesson (2017) att det föreligger DIF (Differential Item Functioning) för uppgifter med flervalsalternativ till pojkars fördel medan uppgifter som kräver ett resonemang är till flickornas fördel. Uppdelningen av flervalsuppgifter och uppgifter med öppna svar sammanfaller delvis med faktorerna *analysera* och *resonera*, men inte fullständigt. Detta antyder att faktorerna delvis kan förklaras av en uppgiftsformat.

5. Diskussion

Resultaten tolkas som att *classification accuracy* och *consistency* (Livingston, 2018; Meyer, 2010) påverkas negativt och med stor variation mellan individuella provtagare genom användning av beläggsmodellen. Slumpens inverkan för hur provbetyg faller ut är betydande. Användningen av beläggsmodellen reducerar alltså reliabiliteten hos provet i religionskunskap och utfallet slår på olika sätt för olika elever, beroende på vilka svar de får höga respektive låga omdömen för.

Proven har god domänteckning då alla delar av kursplanerna är företrädda i proven. Uppgifterna har också varierande svårighetsgrad, ofta i kvalitativt ökande steg. Sammantaget innebär detta i kombination med en dominerande faktor att beläggsmodellen är överflödig och rent av skadlig för säkra provbetyg. Mindre goda prestationer på några uppgifter bör istället kunna kompenseras med

goda prestationer på andra. Goda prestationer på många uppgifter bör rendera högt provbetyg och inte antingen ett högt eller ett lägre beroende på hur några få uppgifter har besvarats och bedömts. Sammantaget är resultaten som rör religionsprovet i överensstämmelse med en studie genomförd på nationella prov i NO-ämnen för årskurs sex (Bach m.fl., 2019). Resultaten är således reproducerbara över några olika typer av nationella prov och skolämnen.

Beläggsmodellens koppling mellan elevers svar på varje enskild uppgift i proven till nivåer i betygsskalan (Bårman m.fl., 2010) kan jämföras med Gustafssons (2008) påpekande om att en konsekvens av Messicks (1989b) utveckling av validitetsbegreppet innebär att *construct validity* inte är tillämpligt för enskilda och från varandra fristående uppgifter. Beläggsmodellen är därmed inte lämplig för summativ användning och bör hanteras försiktigt när det gäller formativ användning, vilket illustreras till exempel genom Habermans test för delskalor (Haberman, 2008). Reliabilitet och svårighetsgrad är inte oväsentligt vid formativt arbete. Ett annat hinder för formativ användning är den sekretess som gäller återanvändning av uppgifter. Beläggsmodellens koppling mellan enskilda elevsvar och nivåbeskrivningar i kunskapskravet bör problematiseras i förhållande till den aspekt av validitet som rör tolkning och användning av resultat och vidare mot de beslut som fattas samt konsekvenser av dessa (Kane, 2013, 2016). Tillämpningen av beläggsmodellen kopplar också tillbaka på resonemang om etiska principer för själva bedömningen som Stern (2017) och Fancourt (2017) för. För vissa elever har beläggsmodellen lett till lägre provbetyg än vad som kan anses vara rimligt medan andra som har lyckats svara bra på rätt uppgifter möjligen har fått för höga provbetyg.

Det finns tecken på att beläggsmodellen har bidragit till förändrade bedömningspraktiker bland lärare (Lidar m.fl., 2017; Lidar m.fl., 2020). Förändrade bedömningspraktiker i enlighet med beläggsmodellen kan leda mot att använda elevers svar från enskilda eller några få uppgifter för summativ och betygssättande bedömning i förhållande till specifika delar av kunskapskravet. Detta får i sin tur möjligen konsekvenser för likvärdighet i betygssättning. Det finns också tecken på förändring i hur formativ bedömning förstås (Lindberg m.fl., 2019) och syn på kunskap och kunskapsutveckling. Dessa aspekter är exempel på områden där mer forskning behövs.

Då det som bedöms genom proven i religionskunskap i huvudsak är endimensionellt kan andra mindre komplicerade och mer robusta metoder för att fastställa provbetyg användas än vad som är fallet idag. Därutöver tar uttolkningen av vårt målrelaterade betygssystem, att provbetygsgränser ska sättas innan proven har genomförts, liten hänsyn till faktisk variation i svårighetsgrad hos annars likartade uppgifter och prov. Det statistiska underlaget från utprovningar innan proven genomförs är både för litet och inte helt representativt för att tillräckligt säkert veta detta. Om hänsyn tas till varierande medelvärde hos proven erhålls en betydligt mer likartad provbetygsfördelning över tid. Med ett sådant tillvägagångssätt är det dock inte möjligt att urskilja trender i kunskapsutvecklingen över tid, men med utgångspunkt i att sådan förändring går relativt långsamt erhålls på detta enkla sätt någorlunda jämförbara provbetyg mellan årskullar nära varandra. Betygens urvalsfunktion till studier på nästa nivå i utbildningssystemet skulle därmed stödjas bättre av de nationella proven. Skolverket har publicerat en rapport med liknande, men betydligt mer långtgående idéer (Skolverket, 2020).

6. Slutsatser

I förhållande till den första forskningsfrågan bidrar de fem eller sex möjliga delskalorna i religionskunskapsprovets i huvudsak inte med tillräcklig reliabel information för att det ska vara lämpligt att använda dem till att påverka provbetygen (Haberman, 2008). I två av de sex versionerna av provet skulle en delskala kunna bidra med ytterligare information, men det är inte försvarligt att i förväg utgå från att det gäller varje aktuell version. Beläggsmodellen tillför därmed inget mervärde jämfört med att sätta provbetygsgränser med utgångspunkt i totalskalan. Istället tillförs betydande inslag av slump för enskilda elevers provbetyg. Provkonstruktörerna undvek i stort sett dessa effekter i provet 2019 utan att ta bort beläggsmodellen. Dock bör påpekas att det vore ett spel för gallerierna att behålla den och det skulle strax genomskådas av lärare, genom egna beräkningar, att villkoren i ett sådant fall är meningslösa. Skolverket har nu beslutat att beläggsmodellen ska tas bort från och med läsåret 2022/2023.[13]

[13] https://youtu.be/Imuwb9oNfBI

Resultaten från denna studie talar för borttagandet av beläggsmodellen och ger stöd åt betygsutredningens förslag att betygssystemet bör byggas upp med kompensatoriska principer (2020).

När det gäller den andra forskningsfrågan ger en bifaktormodell den bästa anpassningen till data av de prövade modellerna, samtidigt som bidragen från bifaktorerna är så små att den enklast tänkbara modellen med endast en faktor trots allt är att föredra, inte minst för att den är enkel att använda vid kravgränssättning. Provet är så att säga tillräckligt endimensionellt för att en sådan modell ska vara tillfyllest.

Den tredje forskningsfrågan är i huvudsak besvarad genom utfallet av modelleringarna till forskningsfråga 2. Det är möjligt och rimligt att använda provets totalskala för att fastställa intervall för de olika provbetygen, helst med någon form av trenduppgifter inom eller parallellt med de olika versionerna för att stabilisera provbetygen mellan åren.

De förhållandevis stora och onödiga slumpeffekter som har tillförts ett annars reliabelt och välutvecklat prov genom beläggsmodellen har genererat felaktiga provbetyg åt fler elever än nödvändigt över ett antal år. Detta leder tillbaka till de frågor som Stern (2017) lyfter om etik i genomförande av nationella prov i förhållande till oönskade konsekvenser för elever (Kane, 2016; Messick, 1989a).

7. Tack

Tack till gruppen för de nationella proven i religionskunskap för data och kommentarer. Tack också till granskarna av artikeln för konstruktiva kommentarer.

Referenser

Bach, F., Frändberg, B., Hagman, M., West, E., & Zetterqvist, A. (2019). Att betygsätta elevers kunskaper - möjligheter och begränsningar. *Nordic Studies in Science Education*, *15*(2, 2), 159–173. https://doi.org/10.5617/nordina.5896

Bartlett, M. S. (1951). The Effect of Standardization on a χ^2 Approximation in Factor Analysis. *Biometrika*, *38*(3/4), 337. https://doi.org/10.2307/2332580

Beaujean, A. A. (2014). *Latent Variable Modeling Using R* (1 uppl.). Routledge.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2 uppl.). The Guilford Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A Tutorial on Hierarchically Structured Constructs. *Journal of Personality*, *80*(4), 796–846. https://doi.org/10.1111/j.1467-6494.2011.00749.x

Bårman, K., Nyström, P., Rask, L., & Åström, M. (2010). *Ämnesproven i biologi, fysik och kemi i årskurs 9*. https://www.skolverket.se/publikationsserier/ovrigt-material/2010/amnesproven-i-biologi-fysik-och-kemi-i-arskurs-9.

Carlstedt, B., & Gustafsson, J.-E. (2005). Construct Validation of the Swedish Scholastic Aptitude Test by Means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology*, *46*(1), 31–42. https://doi.org/10.1111/j.1467-9450.2005.00432.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A Comparison of Bifactor and Second-Order Models of Quality of Life. *Multivariate Behavioral Research*, *41*(2), 189–225. https://doi.org/10.1207/s15327906mbr4102_5

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Cengage Learning.

Dai, S., Wang, X., & Svetina, D. (2022). *subscore: Computing Subscores in Classical Test Theory and Item Response Theory*. https://CRAN.R-project.org/package=subscore

Epskamp, S. (2019). *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*. https://CRAN.R-project.org/package=semPlot

Erickson, G., Åberg-Bengtsson, L., & Gustafsson, J.-E. (2015). Dimensions of Test Performance in English as a Foreign Language in Different European Settings: A Two-Level Confirmatory Factor Analytical Approach. *Educational Research and Evaluation*, *21*(3), 188–208. https://doi.org/10.1080/13803611.2015.1021259

Fancourt, N. (2017). Assessment in Ethics Education: Neoliberalism, Values and Alignment. I O. Franck (Red.), *Assessment in Ethics Education: A Case of National Tests in Religious Education* (s. 163–175). Springer International Publishing. https://doi.org/10.1007/978-3-319-50770-5_8

Franck, O. (Red.). (2017). *Assessment in Ethics Education*. Springer International Publishing. https://doi.org/10.1007/978-3-319-50770-5

Grelsson, G., Bergqvist, G., & Jacobsson, N. (2013). *Nya kursprov i biologi och kemi för gymnasieskolan: bakgrund och resultat*. Umeå universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-80253

Gustafsson, J.-E. (2008). Effects of International Comparative Studies on Educational Quality on the Quality of Educational Research. *European Educational Research Journal*, *7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1

Gustafsson, J.-E., Cliffordsson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan - problem och möjligheter* (s. 120). SNS - Studieförbundet Näringsliv och Samhälle. https://www.sns.se/aktuellt/likvardig-kunskapsbedomning-i-och-av-den-svenska-skolan-problem-och-mojligheter/

Gustafsson, J.-E., & Erickson, G. (2018). Nationella prov i Sverige: Tradition, utmaning, förändring. *Acta Didactica Norge*, *12*(4), Art. 2, 1-20. https://doi.org/10.5617/adno.6434

Haberman, S. J. (2008). When Can Subscores Have Value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. http://www.jstor.org/stable/20172113

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jiang, G., & Mai, Y. (2021). *equaltestMI: Examine Measurement Invariance via Equivalence Testing and Projection Method*. https://CRAN.R-project.org/package=equaltestMI

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). *semTools: Useful tools for structural equation modeling*. https://CRAN.R-project.org/package=semTools

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Lidar, M., Karlberg, M., Almqvist, J., Östman, L., & Lundqvist, E. (2017). Teaching Traditions in Science Teachers' Practices and the Introduction of National Testing. *Scandinavian Journal of Educational Research*, *62*(5), 754–768. https://doi.org/10.1080/00313831.2017.1306802

Lidar, M., Lundqvist, E., Ryder, J., & Östman, L. (2020). The Transformation of Teaching Habits in Relation to the Introduction of Grading and National Testing in Science Education in Sweden. *Research in Science Education*, *50*(1), 151–173. https://doi.org/10.1007/s11165-017-9684-5

Lindberg, V., Eriksson, I., & Pettersson, A. (2019). *Formativ Bedömning: Utmaningar För Undervisningen* (1 uppl.). Natur & Kultur.

Livingston, S. A. (2018). *Test Reliability* (RM-18-01; Research Memorandum, s. 46). Educational Testing Service.

Loehlin, J. C. (2004). *Latent Variable Models an Introduction to Factor, Path, and Structural Equation Analysis* (4. ed.). Mahwah, N.J.: L. Erlbaum Associates.

Messick, S. (1989a). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.2307/1175249

Messick, S. (1989b). Validity. I R. L. Linn (Red.), *Educational Measurement* (3 uppl., s. 13–103). American Council on Education & Macmillan.

Meyer, P. (2010). *Understanding Measurement: Reliability*. Oxford University Press. https://books.google.com?id=x60lzuvWtRYC

Osbeck, C., Franck, O., Lilja, A., & Lindskog, A. (2015). Challenges of Assessment in Ethics – Teachers' reflections when assessing National Tests. *Educare*, *2015*(2), 29.

Osbeck, C., Franck, O., Lilja, A., Sporre, K., & Tykesson, J. (2018). Abilities, knowledge requirements and national tests in RE – The Swedish case as an example in the outcome-focused school and society of today. *Zeitschrift für Pädagogik und Theologie*, *70*(4), 397–408. https://doi.org/10.1515/zpt-2018-0056

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Regeringskansliet, R. och. (2007, maj 3). *Tydliga mål och kunskapskrav i grundskolan - Förslag till nytt mål- och uppföljningssystem* [Text]. Regeringskansliet; Regeringen. https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2007/05/sou-200728/

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Robinson, D., Hayes, A., & Couch, S. (2020). *broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, Articles*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Sinharay, S. (2019). Added Value of Subscores and Hypothesis Testing. *Journal of Educational and Behavioral Statistics*, *44*(1), 25–44. https://doi.org/10.3102/1076998618788862

Skolverket. (2016). *Betygsskalan och betygen B och D*. Skolverket. https://www.skolverket.se/publikationer?id=2953

Skolverket. (2020). *Likvärdiga betyg och meritvärden* [Text]. https://www.skolverket.se/publikationer?id=7582

SOU 2020:43. (2020). *Bygga, bedöma, betygssätta - betyg som bättre motsvarar elevers kunskaper. SOU 2020: Betänkande från Betygsutredningen 2018 (U 2018:03).*

Stern, J. (2017). The Assessment of Ethics and the Ethics of Assessment. I O. Franck (Red.), *Assessment in Ethics Education: A Case of National Tests in Religious Education* (s. 177–191). Springer International Publishing. https://doi.org/10.1007/978-3-319-50770-5_9

Tykesson, J. (2017). Differential Item Functioning in the National Tests in Religious Education in Sweden. I O. Franck (Red.), *Assessment in Ethics Education* (s. 145–162). Springer International Publishing. https://doi.org/10.1007/978-3-319-50770-5_7

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach* (Har/Cdr). Routledge Academic.

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using Bifactor Models to Examine the Predictive Validity of Hierarchical Constructs: Pros, Cons, and Solutions. *Organizational Research Methods*, *24*(3), 530–571. https://doi.org/10.1177/1094428120915522

Åberg-Bengtsson, L., & Erickson, G. (2006). Dimensions of National Test Performance: A Two-Level Approach. *Educational Research and Evaluation*, *12*(5), 469–488. https://doi.org/10.1080/13803610600808115