

Skolverkets systemramverk för nationella prov



Skolverkets systemramverk för nationella prov

Publikationen finns att ladda ner som
kostnadsfri PDF från Skolverkets webbplats:
skolverket.se/publikationer

ISSN: 1652-2508
ISBN: 978-91-7559-313-5

Grafisk produktion: AB Typoform
Skolverket, Stockholm 2017

Innehåll

1. Inledning	4
2. Validitet	5
2.1 Validitetsbegreppet	5
2.2 Precisering av konstruktet och validering	6
2.3 Olika variationskällor vid validering	7
2.4 En modell för hot mot provs validitet	8
2.5 Utformning av validitetsargument och valideringsplan	11
2.6 Specificera syfte och population	12
2.7 Multipla syften	12
2.8 Samla in och dokumentera validitetsinformationen	14
2.9 Negativa konsekvenser	15
2.10 Omprövning av validiteten	16
3. Reliabilitet	17
3.1 Tillräcklig reliabilitet	17
3.2 Variationskällor i provresultat	17
3.3 Mätteoretiska modeller	18
3.4 Minimikrav	21
4. Rättvisa och bias	22
4.1 Hantera rättvisa	22
4.2 Anpassningar av prov	23
5. Provutveckling	24
5.1 Sammansättning av arbets- och referensgrupper	24
5.2 Utveckling av provspecifikationer och modeller för dokumentation av beslut	25
6. Bedömning och rapportering	29
6.1 Bedömning av elevprestationer	29
6.2 Skalor och kravgränssättning	33
6.3 Tolkning av provresultat	37
7. Riktlinjer för provens användning och genomförande	39
7.1 Provtider	39
7.2 Anpassning av prov och genomförande	39
7.3 Provens genomförande, återrapporering och användning	41
7.4 Användning av provresultat	42
8. Referenser	43

1. Inledning

Detta systemramverk syftar till att reglera Skolverkets verksamhet med nationella prov och verksamheten för de lärosäten som på Skolverkets uppdrag utvecklar nationella prov.

Avsikten med systemramverket är att säkerställa högsta möjliga kvalitet i de nationella proven, och högsta möjliga trovärdighet i användning och konsekvenser av resultat på nationella prov i förhållande till de syften¹ som proven har. Ramverkets primära målgrupp är Skolverket och de lärosäten som på Skolverkets uppdrag konstruerar nationella prov. Andra som kan ha intresse av att ta del av systemramverket är forskare och provsystemets användare.

Systemramverket bygger på ett förslag till systemramverk som utarbetats av Gudrun Erickson, Jan-Eric Gustafsson och Peter Nyström vid Göteborgs universitet och som levererades till Skolverket den 30 juni 2017.). En fördjupad bakgrund och diskussion kring frågan om ett gemensamt ramverk för det nationella provsystemet finns att läsa i delrapporten ”Utgångspunkter, argument och ställningstaganden” som Göteborgs universitet redovisade till Skolverket i oktober 2015.² Förslaget till systemramverk och delrapporten har tagits fram i samverkan med grupper av experter som lämnat viktiga bidrag såväl i utvecklingsskedet som i den slutliga granskningen av förslaget till systemramverk. Följande personer har medverkat i olika skeden av arbetet: Ingrid Carlgren, Christian Lundahl, Kristian Ramstedt, Åsa Lindberg Sand, Jörgen Tholin, Jonas Vlachos och Christina Wikström.

Systemramverket (version 1.0) beslutades den 13 december år 2017. Systemramverket kan komma att revideras utifrån förändringar av provsystemet eller politiska beslut. I denna version behandlas till exempel inte frågor om digitalisering av nationella prov och extern bedömning.

Detta systemramverk för nationella prov innehåller:

- beskrivningar av utgångspunkter för provutveckling
- generella riktlinjer samt specifikationer för vad som ska ingå i de provramverk som ska gälla för vart och ett av de nationella proven.

Det som ska preciseras i specifika provramverk för varje prov anges i gröna rutor i texten. Texten innehåller åtta avsnitt, där de första har fokus på grundläggande begrepp och överväganden, nämligen *Validitet*, *Reliabilitet* samt *Rättvisa och bias*. Därefter följer avsnitt om *Provutveckling* samt *Bedömning och rapportering*. Rapporten avslutas med en översiktlig del kring *Riktlinjer för provens användning och genomförande* samt en förteckning av de *Referenser* som görs i texten.

I kapitel 2, 3 och 4 nedan beskrivs utgångspunkter och teoretisk förankring för systemramverket mer i detalj.

1 För närvarande behandlar systemramverket i huvudsak ett av syftena med nationella prov, nämligen nationella prov som stöd för lärares bedömning och betygssättning på individnivå. Om regeringen beslutar om fler syften behöver systemramverket även ta sådana i beaktande i högre grad (se även avsnitt 2.7).

2 Förslag till systemramverk samt rapport dnr: 2015:00077

2. Validitet

Systemramverket har sin utgångspunkt i modern validitetsteori (se till exempel Kane, 2006 och Messick, 1989), vilket innebär fokus på användning i vid bemärkelse, dvs. på de slutsatser, beslut och handlingar som de nationella proven och deras resultat ger upphov till. I systemramverket betonas provens syfte och de implikationer dessa får. Reliabilitet och precision behandlas som en aspekt av validitet, liksom etik och rättvisa. Likaså uppmärksammas provens inflytande och påverkan (*impact*) på tre nivåer – individuell, pedagogisk och systemrelaterad. De kortfattade teoretiska avsnitten konkretiseras med såväl process- som produktrelaterade exempel på krav som ställs i systemramverket. Successiv validering berörs från begreppsliga och operativa synvinklar.

Den moderna validitetsteorin utgör en del av det tvärvetenskapliga fältet ”Educational Assessment” eller ”pedagogisk bedömning”. Inom detta fält har flera olika discipliner och subdiscipliner förts samman för utveckling av teori, metod och praktik avseende bedömning och där såväl kvalitativ som kvantitativ metod utgör nödvändiga hjälpmedel. Här finns sålunda renodlat ämnesteoretiskt innehåll, tillsammans med ämnesdidaktiskt orienterade teoribildningar kring exempelvis hur elever tänker kring olika begrepp och utvecklar kunskap inom olika områden. Dessa discipliner är av stor betydelse vid utveckling av prov- och bedömningssystem, samt givetvis också vid genomförande av bedömningar.

Andra discipliner av fundamental betydelse behandlar frågor kring hur prov med önskvärda egenskaper, till exempel i form av mätsäkerhet, sätts samman. Eftersom prov i allmänhet konstrueras genom sammansättning av olika komponenter i form av uppgifter eller delprov kan detta inte göras på renodlat innehållslig grund, utan måste vägledas av bland annat kvantitativa principer för hur provs egenskaper bestäms av aggregering av provresultat från olika komponenter. Metoder för sådana ändamål har utvecklats inom den psykometriska disciplinen. I samband med användning och uppföljning av prov kommer också en rik arsenal av metoder och teorier utvecklade inom de beteende- och samhällsvetenskapliga fälten till användning.

Sammanfattningsvis bygger pedagogisk bedömning på fyra typer av discipliner: ämnesteori, ämnesdidaktik, psykometri samt beteende- och samhällsvetenskaplig metod. Framgångsrikt provutvecklingsarbete kräver god kännedom om hela detta tvärvetenskapliga fält.

2.1 Validitetsbegreppet

Validitet³ är det mest centrala begreppet inom det pedagogiska bedömningsfältet. Det har utvecklats under en 100-årsperiod och är fortfarande under utveckling (Kane, 2013). Icke desto mindre finns nu en relativt brett omfattad konsensus om innebörden i begreppet. Följande citat från Messick (1989) ger en generell formulering inom ramen för begreppsvaliditetsmodellen:

3 I denna text förekommer relativt sparsamt med referenser, vilket har sin grund i att detta inte huvudsakligen är ett vetenskapligt arbete. I stor utsträckning bygger systemramverket på formuleringar i andra ramverkstexter, som *Standards for educational and psychological testing* (AERA, AEA & NCME, 2014) och de Standards och Guidelines som publicerats av Educational Testing Service (ETS).

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. [kursiveringar i originalet] (Messick, 1989, s. 13)

En första kommentar som kan göras är att denna definition betonar att validitet avser egenskaper hos tolkningar, slutsatser, handlingar eller beslut baserade på provresultatet. Validitet är sålunda inte en egenskap hos ett prov utan avser adekvat tolkning och användning av provresultaten. Enligt detta synsätt är det sålunda felaktigt att tala om ett provs validitet. En andra kommentar är att validitet handlar om i vilken utsträckning slutsatser och beslut har stöd i empiri och teori. Detta kräver sammanställning av information från olika källor och bedömningar av om den samlade kunskapen ger stöd för den föreslagna användningen av provresultaten, vilket innebär att validitetsdiskussionen ständigt kan utvecklas. En tredje kommentar är att om information från ett prov används för olika syften måste separata bedömningar göras av validiteten i var och en av de olika användningarna. En fjärde kommentar till Messicks validitetsdefinition är att den avser såväl provresultat uttryckta i poängform, som i andra former. Definitionen omfattar sålunda användning av såväl kvantitativa som kvalitativa provresultat.

Det måste också betonas att beslutet om ett visst prov ska användas eller ej i allmänhet innebär att hänsyn måste tas även till andra omständigheter och faktorer, som till exempel om provet har negativa konsekvenser som väger tyngre än dess fördelar, eller om det innebär orimligt höga kostnader.

2.2 Precisering av konstruktet och validering

Det första steget i utveckling av ett prov är att precisera den egenskap eller domän som provet ska fänga. I systemramverket kommer termen *konstrukt* att användas för att representera det som provet avser att mäta.

Precisering och avgränsning av konstruktet innebär att dess olika aspekter av innehåll och processer anges och avgränsas från andra konstrukt, och eventuellt även hur det förväntas vara relaterat till andra konstrukt. Denna precisering resulterar i ett preliminärt konceptuellt ramverk för provet, vilket i sin tur ligger till grund för provramverket. I takt med att valideringen av provet successivt utvecklas kan såväl konstrukt, konceptuellt ramverk, provramverk som prov komma att revideras.

En av de mest fundamentala frågorna i valideringen är i vilken utsträckning provet förmår att täcka konstruktet i dess helhet, liksom att inte något ytterligare tillförs. Om provet endast delvis täcker konstruktet innebär detta en validitetsbrist som brukar betecknas "underrepresentation av konstruktet" och om provresultaten influeras av faktorer som inte ingår i konstruktet innebär detta ett hot mot validiteten som brukar betecknas "konstrukt-irrelevant varians". Båda dessa validitetshot kan få allvarliga konsekvenser. En vanlig form av underrepresentation är att provet i för liten utsträckning omfattar sådant innehåll som är svårt att observera och bedöma. Konstrukt-irrelevant varians beror ofta på överutnyttjande av vissa svarsätt, som exempelvis textproduktion, vilket medför att provet tenderar att mäta svarsfärdigheten, snarare än den innehållsliga kompetens som provet är avsett att mäta.

Validering innebär att pröva hållbarheten i den information som provet ger, och sådana studier ger ofta värdefull information, som både kan ge stöd för föreslagna tolkningar och ge grund för att ifrågasätta dem. I allmänhet krävs dock längre serier av undersökningar, baserade på olika informationskällor för att en klar bild av validiteten i användning av provet för olika syften ska kunna framträda. Det är också angeläget att olika informationskällor med olika typer av information används i valideringsstudier eftersom de kan belysa olika validitetsaspekter.

Fem typer av information är särskilt intressanta och relevanta i valideringsstudier: (1) innehåll, (2) svarsprocesser, (3) intern struktur, (4) relationer till andra variabler och (5) konsekvenser. Dessa diskuteras kortfattat nedan.

2.3 Olika variationskällor vid validering

För prov i utbildningssammanhang är såväl relationen mellan innehåll och mål för utbildningen som de anges i styrdokument som provets innehåll och utformning av centralt intresse vid validering av provet. När provet används som stöd för betygssättning är det också av vikt att eleverna har haft möjlighet att lära sig det innehåll som provet tar upp. I sådana sammanhang är det vidare av central betydelse att resultaten för olika grupper av elever inte påverkas av konstrukt-irrelevant varians som har sin grund i uppgifternas och svarssättens utformning (se vidare under "Likvärdighet och bias").

Analyser av hur elever förstår och besvarar olika typer av uppgifter kan vara av stort värde när det gäller att förstå i vilken utsträckning den avsedda tolkningen av provresultaten är rimlig.

I vissa fall anger det konceptuella ramverket att konstruktet består av ett antal separat identifierbara dimensioner, som dock också i allmänhet förväntas vara korrelerade med varandra. I valideringen av provet bör det då undersökas om den förväntade uppdelningen i olika dimensioner också går att återfinna i de observerade elevprestationerna. Detta ska dock inte tolkas så att konstruktet måste delas upp i de olika dimensionerna eftersom det ofta är mer meningsfullt och ändamålsenligt att representera konstruktet med ett samlat provresultat.

Samband mellan ett visst prov och andra prov kan också ge värdefull validitetsinformation. Vissa prov kan förväntas uppvisa höga positiva samband, medan andra inte bör vara högt korrelerade. Om exempelvis ett matematikprov korrelerar högt med ett läsförståelseprov kan detta peka på att matematikprovet ställer otillbörligt höga krav på läsförståelse. Om, å andra sidan, en uppsättning prov avsedda att mäta samma konstrukt endast har låga samband med varandra kan detta peka på validitetsproblem i ett eller flera prov.

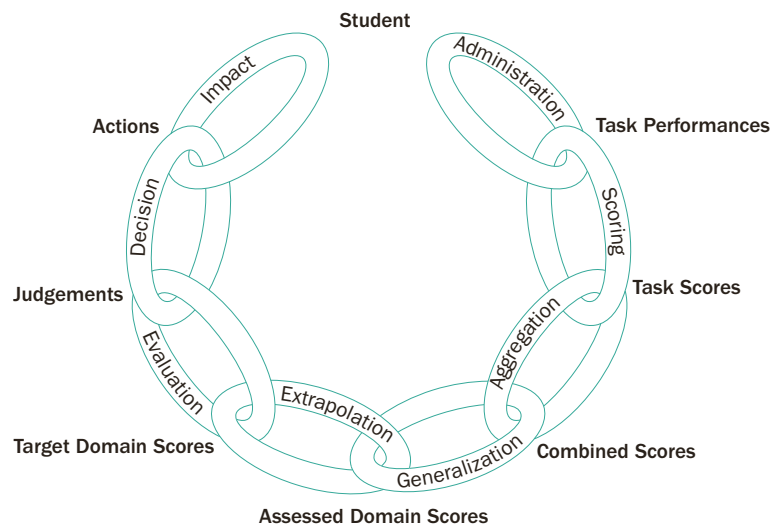
En sista kategori av information kring validitetsproblem avser konsekvenser av provanvändning. Denna "consequential basis" utgör en fundamental del av Messicks begreppsvaliditetsmodell. Konsekvenser kan vara såväl avsedda som icke avsedda, och de kan vara följder av såväl valida som invalida inferenser på grundval av provresultat. Om det exempelvis föreligger en underrepresentation av konstruktet i ett betygstödande nationellt prov, där pojkar har bättre resultat än flickor på de icke-repreresenterade delarna, leder underrepresentationen till en icke valid betygsskillnad till pojkarnas nackdel.

2.4 En modell för hot mot provs validitet

Det finns flera mer preciserade, detaljerade och kompletta modeller över olika hot mot provs validitet (se till exempel Kane, 2013). En modell som förmår att balansera fullständighet och komplexitet med användbarhet och begriplighet har presenterats av Crooks, Kane och Cohen (1996). Modellen visar på hur validering kan systematiseras och hur validitet kan demonstreras genom argumentbaserade överväganden. Eftersom denna modell också på ett förtjänstfullt sätt sätter in grundläggande mätteoretiska begrepp i ett tillämpat sammanhang ges här en kortfattad presentation av modellen.

Modellen (se fig. 1) består av åtta länkade steg, representerade som en kedja, och som avser steg i utveckling eller validering av ett prov. Varje steg är viktigt och den svagaste länken bestämmer kedjans styrka.

Figur 1. A model of educational assessment for use in the validation and planning of assessments.



Källa: Crooks, Kane och Cohen (1996)

Den första länken betecknas *administration*, och avser genomförandet av provet. Crooks et al. (1996) pekar på att detta steg normalt inte ägnas så mycket uppmärksamhet i validitetssammanhang, men de identifierar fyra faktorer som kan vara nog så viktiga. Den första är elevens *motivation*, som om den är för låg innebär att eleven inte anstränger sig i tillräcklig grad. I den svenska diskussionen har detta problem framför allt uppmärksamats i samband med Sveriges deltagande i de internationella undersökningarna, där den enskilda elevens resultat inte har någon individuell betydelse. Motivationsfaktorer kan givetvis också ha betydelse vid genomförande av de nationella proven. Den andra faktorn är provängslan, som snarare är förknippad med hög motivation och som hindrar vissa elever att i provsituationen uppvisa sin faktiska prestationsförmåga. Den tredje faktorn avser förhållanden i provsituationen som kan påverka resultaten negativt, till exempel störande inslag, för kort provtid, eller oklart givna provinstruktioner. Den fjärde faktorn är att elevens förmåga kan underskattas på grund av att eleven inte fullt ut förstått uppgiften. Detta kan bero såväl på

specifika svårigheter för enskilda elever som på oklara uppgiftsformuleringar. En del av dessa hot mot provresultatets validitet går att undvika genom goda förberedelser inför genomförandet av provet, och vissa andra genom individuella anpassningar.

Den andra länken i kedjan är *bedömning* av elevernas svar. I modellen identifieras fem hot mot validiteten som härrör från bedömning. Den första är att bedömningsanvisningarna inte inkluderar alla viktiga aspekter av elevernas prestationer, om till exempel ett prov har fokus på högläsning men enbart uppmärksammar avkodning och läsflyt, inte uttrycksfullhet i läsningen. Om man försöker öka bedömaröverensstämmelse genom att i större utsträckning ange objektiva kriterier kan risken öka för att detta hot blir verklighet. Det andra hotet är att bedömare fäster alltför stort avseende vid vissa sätt att svara, och exempelvis betonar formell korrekthet i skrivna svar även när detta inte är relevant för uppgiften. Ett tredje hot mot validiteten är brister i överensstämmelse mellan olika bedömare, eller hos samma bedömare. Även om det är angeläget att reducera sådan inkonsistens är det viktigt att inte minska den vikt som tillmäts sådana elevprestationer som endast kan bedömas professionellt. Allvaret i detta hot ökas i situationer där bedömare kan ha egenintresse i att göra positiva eller negativa bedömningar av elevernas svar, till exempel då de bedömer svar från identifierbara elever. Anonymisering av proven kan reducera detta hot men skyddar inte mot alltför positiv bedömning av elevprestationer från en viss klass eller skola, om samtliga elever kommer från samma klass eller skola. Det fjärde hotet är att bedömningen kan vara alltför analytisk genom att separata bedömningar krävs av alltför många aspekter, vilket kan leda till att bedömningen av kvaliteten i helheten går förlorad. Det femte hotet är att bedömningen är alltför holistisk, vilket kan leda till förlust av information i synnerhet då omfattande elevarbeten bedöms.

Den tredje länken i kedjan är *aggregering*, vilket innebär att bedömningsresultaten från de olika uppgifterna ska läggas samman till delpoäng eller till en totalpoäng. Här identifieras två validitetshot. Det första är att de uppgifter som läggs samman är alltför heterogena, och att resultaten på olika uppgifter endast har låga samband, vilket leder till att den sammanlagda poängen är heterogen. I denna situation kan det vara fördelaktigt att i stället skapa delpoäng baserade på aggregering av mer homogena uppgifter. Det andra validitetshotet är att olika prestationsaspekter ges otillbörlig betydelse eller vikt. En aggregerad poäng är som mest meningsfull då mer vikt ges åt mer betydelsefulla aspekter av det konstrukt som gäller för provet. Viktning påverkas av antalet uppgifter inom olika områden, av hur bedömning och poängsättning av olika uppgifter gjorts, och av hur stor variationen i poäng är för olika uppgifter.

Den fjärde länken i kedjan är *generalisering*, vilket innebär att dra slutsatser från de observerade resultaten på de använda uppgifterna med de använda bedömningsprocedurerna. Vi kan betrakta uppgifter och procedurer som slumpmässiga urval från en mycket större samling uppgifter och procedurer som lika väl hade kunnat ingå i provet. Vi vill nu bestämma med vilken grad av säkerhet vi kan generalisera från de observerade resultaten till den större samlingen uppgifter och procedurer, som kan kallas vår domän. Resultatens generaliserbarhet avser den korrekthet med vilken de observerade resultaten kan generaliseras till att gälla hela domänen. En individs domänpoäng kan definieras som medeltalet av alla de möjliga prov som kan konstrueras för domänen.

Crooks et al. (1996) identifierar tre validitetshot mot generaliserbarheten. Det första hotet är att de betingelser under vilka provet genomförs ibland i alltför liten utsträckning är standardiserade, till exempel genom att tiden för olika uppgifter tillåts variera, att tiden på dagen tillåts variera, att olika uppgiftsformat kan användas och att olika administratörer av provet arbetar på olika sätt. Genom att standardisera sådana faktorer kan generaliserbarheten ökas, men detta kan också medföra underrepresentation av konstruktet. Om vi exempelvis beslutar att ett prov endast ska genomföras på morgonen, innebär detta en hög grad av standardisering, men också att den variation som kan vara förknippad med provgenomförande vid olika tidpunkter under dagen inte fångas av provresultaten. Det andra validitetshotet mot generaliserbarhet är att inkonsistenta bedömningsprinciper används för olika uppgifter, vilket kan leda till låga korrelationer mellan uppgifterna. Om bedömningsprinciperna görs mer enhetliga kan generaliserbarheten öka, men även detta kan medföra en insnävning av konstruktet. Om syftet är att mäta ett brett konstrukt är en sådan enhetlighet i bedömningsprinciperna rimlig, vilket torde vara fallet för de nationella proven, men inte då uppmärksamheten är fokuserad på enskilda uppgifter. Det tredje, och viktigaste, validitetshotet mot generaliserbarheten är att alltför få uppgifter används. Detta har sin grund i att mätfehlen på olika uppgifter tenderar att balansera ut varandra, och denna balansering stärks om det finns flera uppgifter.

Den femte länken i kedjan är *extrapolering* vilket innebär att de uppgifter som ingår i provet visserligen kan betraktas som ett slumpmässigt urval från den undersökta domänen, men att de i själva verket endast representerar en delmängd av de mål som vi faktiskt är intresserade av. Vi önskar därför göra en extrapolering från den undersökta domänen till måldomänen. Två validitetshot förekommer i detta sammanhang. Det ena är att om den undersökta domänen har studerats med begränsade metoder kan det vara vilseledande att behandla den som ekvivalent med måldomänen. Om exempelvis endast flervalssuppgifter används i den undersökta domänen är det rimligt att förvänta sig resultatskillnader mellan den undersökta domänen och måldomänen, om den senare handlar om att kunna producera egna svar. Det andra hotet är att delar av måldomänen eventuellt inte har undersökts, vilket motsvarar det som Messick betecknar som underrepresentation av konstruktet. Allvaret i detta hot är en funktion av i vilken utsträckning den undersökta domänen avviker från måldomänen.

Den sjätte länken betecknas *evaluerings* och innebär en värdering av innebörden i individens skattade resultat i måldomänen. Crooks et al. (1996) identifierar tre hot mot validiteten i evalueringssteget. Det första är att den person som gör evalueringen har bristande förståelse för tolkning av provinformation och för dess begränsningar. I synnerhet i de fall då personen som evaluerar provresultatet inte själv har konstruerat provet kan missförstånd uppstå. Det andra hotet är att tolkningar av innebörden i konstrukt kan vara svagt underbyggda, och detta gäller i synnerhet abstrakta konstrukt med oklar forskningsunderbyggnad. Ett exempel kan vara konstruktet 'lära att lära' vilket i flerfaldiga försök visat sig svårt att fånga och förstå innebörden av. Det tredje hotet avser olika former av bias i tolkningen av provresultat, och där ett exempel kan vara att tidigare goda (eller svaga) prestationer gör att ett svagt (eller mycket gott) resultat inte tillmäts någon större betydelse.

Den sjunde länken betecknas *beslut*, och avser något av de många typer av beslut som kan följa på ett provresultat. Två hot mot validiteten i fattade beslut

identifieras. Det första hotet utgörs av felaktigt fastställda standards (kravgränser) vilket kan leda till felaktiga provbetyg eller slutsatser om kunskapsnivån och i förlängningen till felaktiga kurs- eller ämnesbetyg. I den typ av beslut där de nationella proven är avsedda att i första hand utgöra underlag, nämligen betygsättning av enskilda elevers kunskaper, kan provresultaten både utöva för stark styrning, så att slutbetyget i alltför hög grad influeras av provbetyget, och för svag styrning, så att slutbetyget i alltför ringa grad influeras av provresultatet. Det andra hotet som anges är dåliga pedagogiska beslut, det vill säga att provresultat ofta ligger till grund för pedagogiska beslut med implikationer på både kort och lång sikt.

Den åttonde och sista länken betecknas *påverkan* ("impact"), och har samma innebörd som Messicks "consequential basis". Crooks et al. (1996) understryker att prov av olika slag ofta har en djupgående påverkan på såväl enskilda elever som på hela undervisningsprocessen. Även om ett prov utvecklats i enlighet med de sju beskrivna punkterna ovan, kan dess negativa påverkan på verksamheten göra att dess validitet ifrågasätts. Två hot mot validiteten som är associerade med denna risk kan identifieras. Det första hotet är att förväntade positiva konsekvenser av provet uteblir. Provsystem innebär betydande insatser av tid och ansträngningar från såväl elever som andra inblandade och i den mån de positiva effekterna av provet inte upplevs motsvara dessa insatser kan negativa effekter förväntas uppstå. Det andra hotet är att provet har en allvarlig negativ påverkan på användare av proven, framför allt eleverna men också lärarna. Som påpekats ovan kan prov bland annat framkalla testängslan, minskad motivation och försämrad självbild, och prov som inte upplevs som rättvisa får inte acceptans. Prov kan också påverka utformningen av undervisningen, på så sätt att provinnehållet får en mer framskjuten plats i undervisningen än vad läroplanen anger.

Med denna generella beskrivning av validitet som grund och utgångspunkt anges nedan huvuddragen i den information som ska finnas i provramverket. En ytterligare mer preciserad information finns i efterföljande avsnitt.

2.5 Utformning av validitetsargument och valideringsplan

De empiriska och teoretiska underlag som finns tillgängliga för att stödja validiteten i de inferenser som görs och åtgärder som vidtas på grundval av provresultaten från nationella prov ska beskrivas. All relevant information ska läggas fram, inklusive sådan som talar mot den föreslagna användningen av provet. I den utsträckning informationen inte finns tillgänglig, anges vilken information som ska samlas in, och formerna för detta.

Utgångspunkten för identifikation och insamling av sådan evidens ska vara dess relevans för att stödja ett validitetsargument. Den typ och omfattning av evidens som krävs för ett visst prov avgörs från fall till fall på grundval av tidigare forskning och professionella bedömningar. Informationen bör avse såväl avsedda utfall (till exempel uppnådda kunskapskrav för ett visst betygssteg) som negativa konsekvenser som följer av ett felaktigt beslut grundat på testet (till exempel ej uppnått betyget E trots att eleven besitter de kunskaper som krävs).

Om argument framförs för att använda ett prov därför att det kan ge positiva indirekta effekter, som exempelvis att provet tydliggör vad eleverna förväntas

lära sig, ska evidens som visar på existensen av sådana indirekta effekter också presenteras. Vederbörlig hänsyn bör också tas till evidens som visar på att provanvändning kan ha negativa effekter på undervisning.

En valideringsplan ska finnas för provet. Systematiska uppföljningar och utvärderingar ska genomföras på det sätt som är preciserat i valideringsplanen.

2.6 Specificera syfte och population

Det eller de konstrukt som provresultaten avser att representera ska preciseras i provramverket. Vidare ska avsedda tolkningar, slutsatser och beslut avseende provtagarna anges, liksom den population provet avser. De sammanhang i vilka provet ska användas ska också anges, liksom hur provet ska administreras och bedömas. I förekommande fall ska provets teoretiska utgångspunkter preciseras tillsammans med hänvisningar till relevant forskning.

2.7 Multipla syften

Detta systemramverk för nationella prov handlar nästan uteslutande om kvalitetsaspekter som rör användning av individuella provresultat som stöd för bedömning av hur väl elever uppfyller kurs- och ämnesplanernas kunskapskrav. Det är vanligt att prov som i första hand designats för sådan användning också används för andra ändamål. Så är fallet i det nuvarande svenska provsystemet, där de nationella proven både ska stödja en likvärdig och rättvis bedömning och betygssättning och ”ge underlag för en analys av i vilken utsträckning kunskapskraven uppfylls på skolnivå, på huvudmannanivå och på nationell nivå”. Det sistnämnda kan beskrivas som att de nationella proven ska kunna användas för utvärdering på olika nivåer.

Newton (2007) utreder olika sätt att se på syften med prov och diskuterar problemen med de grova och ofta oklara syften som anges. Han menar att det inte bara handlar om att provresultat som är anpassade till ett syfte (till exempel betygssättning) kanske inte är särskilt väl anpassade för ett annat (till exempel utvärdering). Det är till och med så att provresultat som är trovärdiga i förhållande till en aspekt av ett syfte (till exempel systemutvärdering på kort sikt) kanske inte är lika trovärdiga i förhållande till en annan aspekt av samma syfte (till exempel systemutvärdering på lång sikt): *We need to convey the complexities of assessment design and fitness-for-purpose; we should not allow those complexities to be over-simplified.* (s. 161). Detta gäller kanske i synnerhet ett utvärderande syfte, som i sig rymmer en mängd olika användningar av provresultat, på kort och lång sikt, på olika nivåer, för olika typer av beslut, m.m. För att nationella prov ska kunna utformas och utvärderas för ett utvärderande syfte är det viktigt att dessa syften preciseras och att komplexiteten i utvärderingssyftet inte döljs i en allmän formulering.

Forskningslitteraturen inom fältet pedagogisk bedömning är mycket tydlig och samstämmig i att det finns stora utmaningar i att använda provresultat för olika syften. Modern validitetsteori är också mycket tydlig med att varje användning av provresultat måste valideras för sig. Samtidigt är det ganska vanligt att provresultat används på mer än ett sätt, till exempel både för att beskriva vad en enskild individ vet och kan göra och för att ge en bild av hur det ser ut för olika grupper. Det finns heller inget som säger att syften är helt oförenliga,

även om det kan vara svårt att uppnå flera syften med samma kvalitet. Ett förhållningssätt är att identifiera ett primärt syfte, som proven optimeras för, och att andra syften får en underordnad roll (Dijkstra, et al., 2012; Newton, 2007). Andra syften kan därmed uppfyllas, men med något lägre grad av säkerhet.

Vad är det då som gör det svårt att optimera prov i förhållande till olika syften? För att resultat från nationella prov ska vara trovärdiga och användbara som omdömen om vad enskilda individer vet och kan göra bör de uppfylla en rad olika krav. Ett krav som är särskilt viktigt i samband med detta syfte handlar om *reliabilitet*. Om det individuella resultatet inte ska bero på slumpmässiga variationer behöver den individuella mätningens medelfel vara litet. Detta kräver i sin tur reliabla prov, vilket betyder att provet måste innehålla många uppgifter som i någon mening mäter samma sak (Wick, 1987, s. 200). För att åstadkomma detta blir det ofta nödvändigt att begränsa antalet områden som mäts i provet. Om områdena är olika och därför vart och ett behöver representeras av ett relativt stort antal uppgifter för att nå reliabilitet, kan det som mäts behöva begränsas.

Om provresultat ska användas för slutsatser på gruppnivå (ett utvärderande syfte) är den enskilda mätningens medelfel inte lika viktigt. Proven behöver ha en rimlig reliabilitet, men genom att resultat från olika elever aggregeras blir det samlade resultat stabilt. Här eftersträvas ofta en bredare domäntäckning, och det är möjligt eftersom alla provtagare inte nödvändigtvis måste svara på samma frågor. Medan den enskilda mätningens medelfel sålunda är av underordnad betydelse då provresultat används för utvärderingsändamål, ställer sådana syften särskilda krav på provresultatens jämförbarhet över tid. Visserligen är stabilitet i provresultaten från ett år till ett annat av betydelse även då proven används för betygsgestödande ändamål, men eftersom huvudsyftet med utvärderingar ofta är att studera förändring är jämförbarhet över tid en nödvändig förutsättning för att dessa syften ska kunna uppfyllas. Skillnader i olika provversioners täckning av olika aspekter av konstruktet är ett vanligt förekommande hot mot validiteten i användning av provet för utvärderande syften.

Det kan också noteras att det sätt på vilket provinformation används för att stödja betygssättningen har avgörande betydelse för vilka kvalitetskrav som bör ställas. I det nuvarande systemet är det individuella provresultatet i fokus, men det skulle också vara möjligt att använda provinformation för styrning av betygssättningen på klass- eller skolnivå. I en sådan modell skulle den enskilda mätningens medelfel vara av mindre betydelse, eftersom de till klass- eller skolnivå aggregerade resultaten får en högre reliabilitet, i enlighet med resonemanget ovan. I en sådan modell skulle sålunda spänningen mellan ett betygsgestödande och ett utvärderande syfte vara mindre.

Sammanfattningsvis är det vanligt att provresultat används på fler än ett sätt, och det kan finnas möjligheter att göra det även om det innebär stora utmaningar:

On the one hand, it does seem theoretically possible to use the same assessment evidence to generate low-level micro-inferences for formative purposes and high-level macro-inferences for other purposes – just as long as the original evidence base is sufficiently comprehensive, structured, and detailed. (Newton, 2010, s. 395)

Samtidigt som det kan vara möjligt att hitta pragmatiska lösningar för att ha multipla syften med nationella prov kan detta förväntas leda till mållkonflikter:

On the other hand, a single assessment system designed to this specification would probably turn out to be both over-engineered for formative purposes (in terms of comprehensiveness and structure) and over-engineered for other purposes too (in terms of detail). (Newton, 2010, s. 395)

Om provresultat ska användas på olika sätt är det nödvändigt att ge validitetsargument för varje användningsområde.

2.8 Samla in och dokumentera validitetsinformationen

Validitetsinformation kan vara konceptuell, empirisk och teoretisk, och indikerar i vilken utsträckning provet uppnår sina syften. Sådan information ska samlas in systematiskt och analyseras och sammanställas i form av ett sammanhängande validitetsargument som ger stöd för korrektheten i de slutsatser som är tänkta att dras och de åtgärder som ska vidtas på grundval av provresultaten. Den samlade evidensen ska enligt resultaten i de genomförda studierna vara tillräcklig för att indikera att provet kan stödja de avsedda tolkningarna av provresultatet för att uppfylla avsedda syften. Studierna och tolkningarna av resultaten i dessa ska granskas i en peer-review process där andra lärosäten är ”kritiska vänner”.

Validitetsargumentet bör presentera den evidens som krävs för en sammanhängande och övertygande argumentation för användning av provet för det avsedda syftet med den avsedda gruppen av provtagare. Det är inte tillräckligt att validitetsargumentet endast utgörs av en sammanställning av den evidens som råkar vara tillgänglig, oavsett dess relevans.

PROVRAMVERK: DOKUMENTERA VALIDITETSINFORMATION

I respektive provramverk presenteras den valideringsplan som gäller för provet i fråga. Planen beskriver när och i vilka steg valideringen genomförs, innehåll och metod för valideringen samt användning och avsedd effekt av de resultat som framkommer. Vidare anges de miniminivåer för provets reliabilitet och konsistens mellan bedömare som gäller för provet. Sammantaget bör informationen innehålla en beskrivning av:

- de procedurer och kriterier som använts för att fastställa provets innehåll, och hur innehållet förhåller sig till det avsedda konstruktet
- i vilken utsträckning expertbedömningar ger stöd för att provet ger information om det avsedda konstruktet
- kvalifikationer hos ämnesexperter, uppgiftskonstruktörer, granskare av uppgifter och prov, och andra personer som är involverade i provutveckling eller validering
- de former under vilka bedömare och experter gjort sina bedömningar, graden av överensstämmelse samt om sam- eller medbedömning använts
- metoder och procedurer som används vid datainsamlingar, representativiteten i de urval av personer som analyserna bygger på, de förhållanden under vilka uppgifterna samlades in, resultat totalt och för undergrupper av populationen samt precisionen i den redovisade kvantitativa informationen
- statistiska samband mellan provets delar
- argument och belägg för eventuella föreslagna tolkningar av svar på enskilda uppgifter, delmängder av uppgifter samt del- och/eller profilsummor
- relationer mellan provpoäng eller andra provresultat och externa variabler, inklusive motiven för val av externa variabler, deras egenskaper, och relationerna mellan dem.

2.9 Negativa konsekvenser

Om användningen av ett prov har oavsiktliga negativa konsekvenser ska validitetsbeläggen prövas för att avgöra om de negativa konsekvenserna härrör från konstrukt-irrelevant varians. Om de gör det ska provet revideras för att i möjligaste mån minska den konstrukt-irrelevanta variansen. Exempelvis visar en omfattande forskning att kvinnor har bättre skrivförmåga än män. Om då uppgifter som kräver mer omfattande skriftlig produktion används även då detta inte är en central del av konstruktet, medför sådana uppgifter konstrukt-irrelevant varians som bidrar till uppkomst av validitetshotande könsskillnader i provresultat.

Oavsiktliga negativa konsekvenser behöver inte nödvändigtvis göra användning av ett prov ogiltig. Det är dock nödvändigt att undersöka huruvida de oavsiktliga konsekvenserna kan kopplas till faktorer som är konstrukt-irrelevanta eller till underrepresentation av konstruktet, och om så är fallet måste korrigerande åtgärder vidtas. Sådana bör också vidtas då det oavsett orsaken till de negativa konsekvenserna är möjligt att göra detta utan att försämra validiteten.

2.10 Omprövning av validiteten

Om relevanta faktorer förändras, ska beläggen prövas för att se om validitetsargumentet för provet fortfarande uppfyller det avsedda ändamålet och stöder den avsedda tolkningen av provresultaten för den avsedda gruppen. Denna prövning innebär också insamling av den nya information som krävs.

De faktorer som kan föranleda sådan omprövning inkluderar till exempel väsentliga ändringar i den teknik som används för att administrera eller bedöma provet, avsedda ändamål med provet, den avsedda tolkningen av provresultat, provets innehåll eller populationen av provdeltagare.

Det finns ingen tidsgräns inom vilken omprövningen av validitetsargumentet ska ske. Däremot bör ett lämpligt intervall bestämmas mellan återkommande översyner och en logisk grund för det intervall som valts bör anges.

3. Reliabilitet

Ett provs reliabilitet avser i vilken utsträckning provresultaten kan generaliseras till att gälla andra versioner av provet, andra bedömare av elevprestationerna, och eventuellt även till andra provtillfällen. Som framgår av beskrivningen av validitetsbegreppet ovan ingår reliabilitetsbegreppet som en del av detta, men det senare begreppets tekniska karaktär motiverar att det även behandlas under en egen rubrik.

De olika former av information om reliabilitet (till exempel reliabilitets- eller generaliserbarhetskoefficient, informationsfunktion, och betingade medelfel) ska vara adekvata med tanke på provets avsedda användning, den avsedda populationen, och de psykometriska modeller som används för att konstruera provresultat.

3.1 Tillräcklig reliabilitet

Den reliabilitetsnivå som krävs för ett visst prov kan endast avgöras genom en professionell bedömning, som beaktar provets syfte och de konsekvenser ett felaktigt beslut kan få. För prov som används som underlag för beslut av stor betydelse för enskilda elever är en hög reliabilitet av stor vikt.

3.2 Variationskällor i provresultat

De metoder som används för att bestämma provets reliabilitet ska vara adekvata för det aktuella provet och för provets avsedda syfte. De variationskällor som påverkar provresultaten ska identifieras, och de metoder som används för att bestämma reliabiliteten ska kunna fånga upp dessa variationskällor. Detta innebär att olika typer av prov kan kräva olika metoder för reliabilitetsbestämning. Följande variationskällor är vanligt förekommande:

- Variation som en följd av sammansättningen av uppgifter. Om allt annat är lika är ett prov med många uppgifter mer reliabelt än ett prov med få uppgifter. Reliabilitetseffekter av uppgiftsvariation mäts ofta med internkonsistensmått (till exempel Cronbachs α).
- Variation som en följd av provtillfälle. Upprepning av ett och samma prov ger i allmänhet inte perfekt samband mellan provresultaten för en grupp elever. Reliabilitetseffekter av tillfällesvariation mäts ofta med test retest-metodik.
- Variation som en följd av bedömning av elevprestationer. Olika bedömare gör inte alltid samma bedömning av en och samma elevprestation, och samma bedömare gör inte heller alltid samma bedömning av en och samma elevprestation vid olika tillfällen. Reliabilitetseffekter av bedömarvariation mäts ofta med hjälp av mått på procentuell eller annan form av överensstämmelse mellan bedömningar.

Olika typer av prov influeras ofta av flera olika variationskällor och det är angeläget att de valda måtten förmår att fånga alla variationskällor:

- För prov med mer omfattande elevprestationer (till exempel uppsatser) ska mått som beskriver reliabiliteten i bedömningarna beräknas, liksom även mått som beskriver reliabiliteten i hela provprocessen, inklusive val av uppgifter som föreläggs provtagarna och val av bedömare av elevernas svar och lösningar.
- För prov som redovisar resultat för flera olika kunskapsdomäner eller förmågor ska metoder för reliabilitetsbestämning användas som ger utrymme för provtagare att ha olika resultatnivåer inom de olika domänerna eller förmågorna.
- För prov och delprov som används för att klassificera elevprestationer i olika kategorier (till exempel godkänt och icke godkänt) på grundval av deras provresultat bör mått beräknas som anger konsistensen över olika provversioner för dessa klassifikationer.
- För alla prov gäller att reliabilitetsmått ska beräknas som är adekvata för den aggregationsnivå som resultaten ska rapporteras på (till exempel enskild elev, grupp-/klassnivå och skolnivå).

3.3 Mätteoretiska modeller

I detta avsnitt diskuteras vilka olika former av reliabilitetsinformation som är viktig att ange för de nationella proven, och de olika modeller och metoder som kan användas. I provramverket för respektive prov ska såväl val av metoder som kravnivåer preciseras.

Som redan påpekats är hög reliabilitet av stor vikt då provinformationen används som underlag för beslut av stor betydelse för enskilda elever. Av tradition brukar som ett minimikrav anges att reliabiliteten, fastställd med ett mått på intern konsistens, ska vara minst 0,90 i sådana situationer, där kravet avser det underlag på vilket beslut fattas (till exempel helt prov eller delprov). Denna tumregel är dock mycket grov, och den behöver kvalificeras med hänsyn till specifika syften och användningar.

De nationella provresultaten används ofta i olika former av sambandsstudier, där provresultat relateras till exempelvis bakgrundsvariabler för eleverna, resultat på andra samtidigt givna prov eller betyg, eller framtida utfallsvariabler. I sådana sambandsanalyser används i allmänhet konventionella statistiska metoder som korrelations- eller regressionsanalys. En enkel princip är att i korrelationsanalyser påverkas det beräknade sambandet negativt av reliabilitetsbrister i den ena eller båda av de analyserade variablerna och där underskattningen av sambandet är direkt proportionell mot reliabilitetsbristerna i båda variablerna. Om exempelvis det faktiska sambandet mellan två variabler är 1,0 och de två variablerna mäts med prov som båda har reliabiliteten 0,7 blir inte det observerade sambandet högre än 0,7. En underskattning av ett faktiskt samband i denna utsträckning framstår i många sammanhang som oacceptabel, men med en reliabilitet om 0,90 skulle avvikelsen förmodligen i många fall kunna betraktas som tolerabel.

Då man beräknat reliabiliteten är det också möjligt att bestämma medelfelet för den enskilda mätningen, vilket är standardavvikelsen i den tänkta fördelning av provresultat som skulle erhållas om en individ upprepade ett visst prov

ett oändligt antal gånger. Medelfelet kan sedan användas för att bestämma ett konfidensintervall på till exempel 95 %-nivån, vilket anger det poängintervall inom vilket en elevs provresultat kan förväntas falla i 95 procent av mätningarna (Skolverket, 2015, s 11). Sådan information ger en konkret bild av omfattningen av osäkerheten i mätningen, uttryckt på den skala som provresultaten redovisas i. Som visas i Skolverket (2015, s 11–19) kan dock även ett prov med en reliabilitet nära 0,90 ha ett brett poängintervall för det ”sanna” värdet (true score). Reliabilitetsmått i form av internkonsistenskoefficienter ger sålunda endast mycket grov information, och måste kompletteras med analyser av den enskilda mätningens medelfel.

Ett problem med denna beräkning och tolkning av reliabilitetsinformationen är att den bygger på antagandet att medelfelet för mätningen är lika stor för alla nivåer på skalan. Detta är ett av den klassiska mätlärens grundläggande antaganden, och det finns situationer när det är rimligt att göra ett sådant antagande. De situationerna är dock sällsynta, och i normalfallet är det mer rimligt att anta att medelfelet storlek varierar med nivån på elevens prestation.

För att kunna beräkna medelfel som bygger på det mer rimliga antagandet att dessa varierar som en funktion av prestationsnivån kan vi modifiera den klassiska mätteorin och föra in vad som kallas betingade medelfel, eller så kan vi lämna den klassiska mätläran och i stället gå över till den moderna mätläran.

3.3.1 Betingade medelfel

Formeln för att beräkna den enskilda mätningens medelfel kan relativt enkelt modifieras så att betingade medelfel erhålls i stället (se Skolverket, 2015, s 21). Generellt visar beräkning av betingade medelfel att dessa blir större än det obetingade medelfelet för värden nära medelvärdet, medan de blir mindre för värden nära skalans ändpunkter. Det obetingade medelfelet kan sålunda ses som ett medelvärde av de betingade medelfelen. Det finns starka rekommendationer i olika *Standards* att använda betingade medelfel vid bestämning av mätfels omfattning.

Skolverket (2015) presenterar resultat av ”classification accuracy” i ett genomfört nationellt prov i fysik årskurs 9 (18 uppgifter, maxpoäng 38) genom att undersöka hur stor andel av eleverna som på grundval av sina provresultat hamnar i rätt betygskategori vid tillämpning av de fastställda betygsgränserna. Resultaten visade att 77 procent av eleverna fick ”rätt” betyg, medan 10 procent fick för låga betyg, och 13 procent fick för höga betyg. Alla avvikelser från det ”rätta” betyget var uttryck för slump effekter, eftersom analysen inte tog någon hänsyn till systematiska fel. Rapporten redovisar också intressanta analyser av i vilken utsträckning det var möjligt att utvinna ytterligare information genom att summera resultaten på uppgifterna till en separat G-skala och en separat VG/MVG-skala och att sedan kombinera de två skalorna vid tilldelning av provbetyg. Resultaten visade att detta inte medförde någon förbättring av resultaten, trots en avsevärt mer komplex procedur.

I rapporten understryks att effekterna av de slumpmässiga felen är anmärkningsvärt stora, och att den tekniska rapporteringen bör redovisa sådan information som gör det möjligt att bedöma provens mätmässiga kvalitet.

3.3.2 Reliabilitet i modern mätlära

Den moderna mätläran (eller *item-response theory*, *IRT*) omfattar en stor familj av statistiska modeller som har det gemensamt att man söker bestämma egenskaper hos provuppgifter (till exempel svårighetsgrad och diskriminationsförmåga) och personer (till exempel förmåga). Genom att skatta parametrar i en explicit statistisk modell över sannolikheten för ett korrekt provsvar är det möjligt att identifiera invarianta uppgiftsegenskaper, vilka i sin tur gör det möjligt att på ett flexibelt sätt konstruera och optimera prov. Ett värdefullt hjälpmedel i detta sammanhang är den s.k. informationsfunktionen, vilken är nära relaterad till den enskilda mätningens medelfel.

Informationsfunktionen uttrycker den mängd information som besvarande av en uppgift ger och denna är som störst då sannolikheten för ett korrekt svar är 0,50, och minskar sedan successivt då sannolikheten går mot 0 eller mot 1. Testinformationsfunktionen (TF) är summan av informationsfunktionerna för samtliga de uppgifter som ingår i ett prov, och inversen av TF är medelfelet. TF och medelfel varierar som en funktion av förmåga, så lägst medelfel får man för personer som i medeltal har sannolikheten 0,50 att besvara uppgifterna i provet rätt. Medelfelet ökar sedan för personer med lägre eller högre förmåga.

Inom den klassiska mätläran är sålunda medelfelet lågt för elever som har de högsta och de lägsta resultaten, medan det inom den moderna mätläran är för dessa grupper av elever som vi observerar de största medelfelen. Denna paradox förklaras av att den skala som används inom den klassiska mätläran är en poängskala, som i allmänhet sträcker sig från 0 till maxpoängen på provet, medan skalan inom den moderna mätläran sträcker sig från minus oändligheten till plus oändligheten. Inom den klassiska mätläran uttrycks medelfelet i termer av poängskalan och när det vid skalans ändpunkter inte finns något utrymme för variation minskar medelfelet. Inom den moderna mätläran finns det utrymme för variation i medelfelet över hela skalan men för de provtagare som endast besvarar ett fåtal uppgifter har vi mycket begränsad information om var på skalan de egentligen hör hemma, vilket kommer till uttryck i form av ett stort medelfel. Samma sak gäller för de elever som svarar rätt på praktiskt taget samtliga uppgifter.

Beroendet av en viss skala, knuten till en viss uppsättning uppgifter, är en av de svagheter som vidlåder den klassiska mätläran, medan det relativa oberoendet av specifika uppgifter bidrar till den moderna mätlärans förmåga att lösa många av de mätmässiga utmaningar som ett nationellt provsystem erbjuder. En av dessa är att bestämma omfattningen av mätfel för elever på olika prestationsnivåer, men den erbjuder också tekniker som är användbara i samband med utprövning av uppgifter och sammansättning av prov, liksom för att ekvivalera olika versioner av ett prov.

Den moderna mätläran är i första hand ändamålsenlig för prov som omfattar ett större antal uppgifter, vilka kan vara dikotoma eller omfatta flera ordnade kategorier. För omfattande och komplexa uppgifter erbjuder den moderna mätläran inte lika kraftfulla verktyg, men här kan kombinationer av klassiska och moderna tekniker vara värdefulla.

Av de skäl som framförts ovan ska de mätmässiga frågorna i första hand angripas med tekniker hämtade från den moderna mätläran men då dessa inte är tillämpliga eller ändamålsenliga kan klassiska tekniker brukas.

3.4 Minimikrav

Nedan anges de minimikrav på mätprecision som ska gälla för de betygsstödjande nationella proven. Kraven avser normalt det totala provresultatet, men i den mån betygsstödet endast avser delar av provet gäller kraven för dessa delar. Metoder som innebär separata poängbestämningar för olika betygssteg, till exempel olika former av kvalitetspoäng, ska inte användas om det inte finns väl dokumenterat stöd för att detta medför förbättrad mätsäkerhet.

- *Reliabilitet*: Minst 0,90, beräknad med Cronbachs α .
- *Medelfel*: Beräknas från testinformationsfunktionen. Inga numeriska minimikrav, men kommentarer ges om provets förmåga att ge pålitlig information för olika prestationsnivåer.
- *Klassificeringskorrekthet*: Korrektheten i klassificeringen i provbetyg beräknas dels för samtliga betygssteg, dels för betyget F kontra övriga betygssteg. Inga numeriska minimikrav, men kommentarer ges om provets förmåga att stödja korrekt klassificering.

För performansuppgifter med komplex elevproduktion som kräver bedömning ska interbedömarreliabilitet beräknas, varvid hänsyn ska tas såväl till det relativa felet (dvs. skillnader i rangordning av elever) som till det absoluta felet (dvs. skillnader i stränghet mellan bedömare) enligt generaliserbarhetsteorins terminologi. Minimikrav för interbedömarreliabiliteten preciseras i provramverket, men bör inte understiga 0,6.

4. Rättvisa och bias

För betygsstödjande prov är det viktigt att provresultaten för elever med bland annat olika kön, social bakgrund, migrationsbakgrund och funktionsnedsättning är rättvisa och inte uttrycker något annat än elevernas visade kunskap. Hot mot denna typ av rättvisa kan beskrivas i termer av systematiska mätfel (konstrukt-irrelevant varians) och är därför ett hot mot provets validitet. Det finns en lång rad källor till sådana validitetsproblem, som provgenomförandet, provets innehåll, provuppgifternas och svarssättens utformning, bedömningen av elevprestationer och skalkonstruktion. Förändringar av provgenomförande kan därför behöva göras, och genom hela provkonstruktionsprocessen måste skyddsåtgärder mot de olika formerna av validitetshot vidtas. Prov måste kunna fungera för den mångfald av populationer för vilka det används, så att resultaten är jämförbara och rättvisa, utan att påverkas av skillnader i egenskaper som inte är relevanta för den avsedda användningen.

Det finns många, delvis motstridiga, definitioner av rättvisa i litteraturen. En grundläggande definition av rättvisa är den utsträckning i vilken inferenser som görs på grundval av provresultat är valida för olika grupper av provtagare.

Det bästa sättet att förbättra rättvisa för alla provtagare är att i möjligaste grad minimera inflytandet från konstrukt-irrelevant varians. Det är inte möjligt att göra separata studier av rättvisa för alla undergrupper inom populationen, men det är viktigt att undersöka rättvisa för de grupper som enligt erfarenhet och tidigare forskning riskerar att drabbas av konstrukt-irrelevant påverkan på sina provresultat. Det handlar ofta om grupper som diskriminerats på grundval av faktorer som etnicitet, funktionsnedsättning, kön eller modersmål.

4.1 Hantera rättvisa

Prov ska designas, utvecklas, administreras och bedömas så att de prövar det avsedda konstruktet och minimerar inflytandet från konstrukt-irrelevanta egenskaper hos provtagarna. För ett nyutvecklat prov ska det finnas en rättviseplan för hur frågor kring rättvisa ska behandlas i design, utveckling, administrering och bedömning av provet. För ett existerande prov ska de åtgärder som vidtagits för att hantera rättvisa dokumenteras, tillsammans med dokumentation av framtida rättviseplaner.

Alla provtagare ska behandlas enligt samma principer under provadministration och bedömning av elevprestationer. Av dokumentationen av rättviseplanen ska framgå hur förväntade problem med rättvisa har hanterats eller ska hanteras. Anpassningar av prov för att göra dem tillgängliga för elever med funktionsnedsättningar diskuteras i avsnitt 4.2. En version av rättviseplanen bör finnas tillgänglig för extern spridning.

Gruppskillnader i prestationer innebär inte nödvändigtvis att ett prov brister i rättvisa, men om skillnaderna är stora nog för att få praktiska konsekvenser bör undersökningar genomföras för att säkerställa att skillnaderna inte har sin grund i konstrukt-irrelevanta faktorer, till exempel alltför homogena provformat.

PROVRAMVERK: HANTERA RÄTTVISA

Vad som bör ingå i dokumentationen av rättviseplanen beror på provets karaktär. Om det bedöms relevant för provet, och om det är möjligt att få tillgång till data, bör information om följande inkluderas:

- Granskningar som genomförs för att undersöka rättvisa, inklusive information om bedömarnas kvalifikationer.
- Hur lämpligt provmaterialet är för personer som ingår i undersökta grupper.
- Bedömningar av lingvistiska och läs- och skrivmässiga krav för att säkerställa att dessa inte är högre än vad som är nödvändigt för att uppnå syftet med provet.
- Icke avsedda negativa konsekvenser av användning av provet för olika grupper.
- Resultat från kvantitativa analyser avseende skillnader i provets funktion för olika grupper av elever.
- Granskning av procedurer som använts för att bedöma komplex elevproduktion.
- Gruppskillnader med avseende på användning av olika provtagningsstrategier, tillgång till provträning eller *speededness*.
- Effekter av olika omfattning av erfarenhet av olika provsituationer.

4.2 Anpassningar av prov

Genomförandet av nationella prov ska anpassas så att proven så långt som möjligt görs tillgängliga för elever med funktionsnedsättning.

Anpassningarna ska utformas så att provet i största möjliga utsträckning prövar det avsedda konstruktet, snarare än irrelevanta variationskällor. I mån av tillgång till tillräckligt stora urval bör lärosätet som utvecklar prov genomföra empiriska studier av vanligt förekommande anpassningar. Dessa bör baseras på kunskap om effekter av både funktionsnedsättningar och god provpraktik. Personer med specialpedagogisk kompetens bör ingå i arbetet med utveckling av proven, såväl generellt som med avseende på anpassningar av skilda slag.

Anpassningar ska medges endast då det finns behov av sådana. De som har ansvarat för utvecklingen av provet, ska också förse de som ansvarar för provgenomförande med information om hur anpassningar kan genomföras.

5. Provutveckling

I detta avsnitt av systemramverket beskrivs hur prov ska utvecklas i enlighet med välplanerade och dokumenterade procedurer och under medverkan av personer med relevant kompetens. Kompetens innebär här dels kunskaper inom det tvärvetenskapliga fält som tidigare nämnts (ämnesteori, ämnesdidaktik, psykometri och beteende- och samhällsvetenskaplig metod), dels gedigen erfarenhet av undervisning inom det ämnesområde som är aktuellt samt med elever i relevanta åldrar. Fokus för arbetet är att skapa prov som stödjer en rättvis och likvärdig, transparent och tillförlitlig bedömning, som leder till slutsatser, beslut och åtgärder som är giltiga för sitt syfte och för den avsedda målgruppen.

God provutveckling⁴ förutsätter tydliga specifikationer, successiv kvalitetskontroll, empirisk information om enskilda uppgifters kvalitet samt utvärdering av givna prov, när det gäller såväl analyser av resultat i relation till utprövningar som uppfattningar uttryckta kring provet. Detta kräver samverkan med olika kategorier av experter när det gäller till exempel ämne, undervisning, mätning, genomförande och bedömning, och det inkluderar även användarledet, framför allt lärare och elever.

Varje prov ska åtföljas av två typer av rapporter. Första gången ett nationellt prov har utformats och genomförts ska en initial utvecklingsrapport sammanställas av lärosätet som ansvarar för utveckling av provet. Denna rapport ska innehålla en ingående analys av alla systemramverkets aspekter. Till varje följande prov som utformas utifrån samma provramverk ska en underhållsrapport av mindre omfattning sammanställas.

I det följande ges kortfattade riktlinjer för provutvecklingsarbetet från förarbete fram till dess att uppgifter satts samman till ett färdigt prov. För varje nationellt prov ska ett provramverk, inklusive konkreta specifikationer, utvecklas på basis av dessa riktlinjer. Provramverket ska utgå från de överväganden kring validitet, reliabilitet och rättvisa som görs i första delen av detta systemramverk samt i tillämpliga delar även inkludera det som uttrycks i avsnitt 6, Bedömning och rapportering.

5.1 Sammansättning av arbets- och referensgrupper

Utvecklingen av prov förutsätter samverkan med olika kategorier av experter. För detta ändamål sätts grupper samman för medverkan i olika skeden av arbetet. Exempel på fokus för dessa grupper är konstruktion av uppgifter, granskning av uppgifter i olika delar av utvecklingsarbetet, sammansättning av utprövningsversioner och skarpa prov, förslag till betygsgränser (kravgränssättning) och urval av autentiska nivåexempel som kommenteras i relation till styrdokumentet (*benchmarking*). Det är centralt med bredd när det gäller relevanta kompetenser och erfarenheter (se ovan) och aktiva lärare ska utgöra en betydande del av de grupper som tillsätts.

4 Det finns en rikhaltig litteratur kring provutveckling, såväl av generell som i huvudsak ämnes-specifik art. Litteratur av detta slag är dock alltid till viss del kontextuell och rekommendationer och exempel därför inte automatiskt överförbara till specifika sammanhang. Exempel på mera generella, delvis klassiska referenser är Downing & Haladyna (2006), Ebel (1951) samt Haladyna (1997).

PROVRAMVERK: REFERENSGRUPPER

I provramverket ska de olika grupper som medverkar i provutvecklingens olika faser beskrivas tillsammans med principer för gruppssammansättning. Provråm-
verket ska ange varje grupps funktion och mandat och ange principer för hur
grupperna sätts samman med avseende på gruppmedlemmarnas antal, roll,
kompetens och erfarenhet. Här ska också rutiner för förnyelse av gruppernas
sammansättning beskrivas.

5.2 Utveckling av provspecifikationer och modeller för dokumentation av beslut

För att precisera och standardisera prov samt möjliggöra systematisk uppföljning
av provens funktionalitet och stabilitet, krävs en provspecifikation för varje prov.
Denna del av provramverket beskriver i så konkreta och precisa termer som
möjligt såväl process som produkt, dvs. både provutvecklingsprocessen med dess
olika stadier och produkten/provet (preciserat i punkterna nedan). Provet ska
följa specifikationen så nära som möjligt, och förändringar i provet ska föregås
av välgrundade och explicita justeringar av specifikationen. Provspeçifikationen
ska vara så stabil som möjligt över tid, för att skapa möjligheter till jämförbara
resultat. Provspeçifikation bör utvärderas och justeras vid behov med jämna
mellanrum, förslagsvis vart tredje år.

PROVRAMVERK: PROVSPECIFIKATION

Specifikationen för enskilda prov ska innehålla teoretiskt förankrad, adekvat
och konkret information om:

- provets syfte och målgrupp
- provets kunskapsdomän och konstrukt
- provets format
- provets struktur, omfattning, uppgiftsformat, tidsåtgång och sekvensering
- steg i processen att utveckla provet: konstruktion, granskning och utprövning
- avsedda egenskaper för enskilda uppgifter och för provet i sin helhet
- slutgiltig provsammansättning samt principer för process och produkt.

Nedan ges konkreta kommentarer kring vad som ska ingå i provspecifikationen
i förhållande till var och en av punkterna ovan.

5.2.1 Provets syfte och målgrupp

Utgångspunkten för provutvecklingsprocessen är att syftet för det aktuella
provet är klart, liksom de tänkta användarna och användningarna, den avsedda
kunskapsdomänen samt den grupp vars kunskaper ska bedömas.

PROVRAMVERK: PROVETS SYFTE OCH MÅLGRUPP

I provramverket ska nationellt fastlagda syften med det aktuella provet preciseras och konkretiseras i relation till kunskapsdomänen och den målgrupp som provet gäller.

5.2.2 Provets kunskapsdomän och konstrukt

Provets övergripande kunskapsdomän beskrivs med hänvisning till styrdokument, relevant forskning, nationella och internationella exempel samt erfarenheter av lärande, undervisning och bedömning. Mot bakgrund av detta preciseras det aktuella materialets konstrukt med avseende på innehåll, typ och omfång av det som inkluderas i provet.

PROVRAMVERK: KUNSKAPSDOMÄN OCH KONSTRUKT

I provramverkets provspecifikation ska det eller de konstrukt som provet avser att pröva preciseras. Preciseringsen ska behandla kunskapsdomänen i stort, vilket innebär samstämmighet mellan alla prov i samma ämne. Den ska också vara konkret i relation till det aktuella materialet. I provramverket ska även de procedurer och kriterier som använts för att avgöra provets innehåll beskrivas. Här ska grad och art av domäntäckning anges (vad i styrdokumenten som provas respektive inte provas, i enstaka prov eller mera permanent), tillsammans med beskrivningar av hur det som provas förhåller sig till domänen som sådan och till andra aspekter av domänen som provas. Slutligen ska provramverket ange vilken typ av expertis som används vid precisering av provets konstrukt.

5.2.3 Provets format

Provets format, dvs. uppgiftstyper och svarsformer, väljs och preciseras på basis av överväganden kring validitet, reliabilitet och rättvisa, liksom kring ändamålsenlighet för sitt syfte och i relation till sin kontext och målgrupp. I detta beaktas från skilda utgångspunkter för- och nackdelar med olika uppgiftstyper och svarsformat, till exempel flervalsuppgifter, uppgifter med korta svarsformat och så kallade autentiska uppgifter.

PROVRAMVERK: PROVETS FORMAT

I den del av provramverket som kallas provspecifikation ska de uppgiftstyper och svarsformat som används i provet preciseras och förankras i teoretiska överväganden och överväganden om genomförbarhet. Detta gäller i synnerhet balansen mellan flervalsuppgifter och uppgifter som kräver elevproducerade svar av olika längd.

5.2.4 Provets struktur, omfattning, tidsåtgång och sekvensering

Överväganden kring innehåll och format förankras i resonemang om grundläggande principer kring validitet, reliabilitet och rättvisa och relateras till frågor om provets struktur, omfattning, tidsåtgång och sekvensering av uppgifter.

PROVRAMVERK: PROVETS STRUKTUR

I provramverket ska provets struktur beskrivas och motiveras vad avser omfattning i innehåll och tid (delar respektive helhet), samt eventuell uppdelning i delprov. I relation till detta ska principer för de beslut som ligger till grund för sekvenseringen av uppgifter anges.

5.2.5 Steg i processen att utveckla provet: konstruktion, granskning och utprovning

Provutveckling är en iterativ, dvs. upprepad process där justeringar och revideringar görs på grundval av teoretiska överväganden, successiva granskningar, analyser av empiriska data från utprovningar, erfarenheter från autentisk användning av tidigare uppgifter samt en kontinuerlig dialog med användare.

PROVRAMVERK: PROCESSEN FÖR PROVUTVECKLING

I provramverket ska processen för konstruktion av provuppgifter beskrivas och de olika stegen i processen motiveras. Detta gäller särskilt frågor kring metoder att stärka konsistent bedömning och att minimera risken för systematiska fel och ojämn domäntäckning. Vidare ska provramverket behandla provets möjliga påverkan, såväl avsedd som icke avsedd, på elevers möjligheter att visa sina kunskaper och lärares möjligheter att dra adekvata slutsatser av resultaten. Principer för poängsättning ska beskrivas och motiveras begreppsligt och empiriskt; detta gäller även eventuella viktningar av poäng (se vidare avsnitt 6). Instruktioner till uppgifter utgör en viktig del såväl av konstruktion som av utprovning.

Provråmverket ska också beskriva hur provuppgifterna granskas och provas ut successivt under utvecklingsprocessen samt hur resultaten från dessa steg i processen påverkar det fortsatta arbetet med provet.

Vidare ska provramverket precisera och motivera principer för urval av uppgifter för den slutgiltiga utprovning som ska ligga till grund för sammansättningen till ett färdigt prov. Detta gäller bland annat grad av representativitet i urvalet samt antal medverkande skolor respektive individer.

Beskrivningen ska också behandla sammansättningen av utprovningsprov, användning av ankaruppgifter (typ och antal) samt granskningen av de provmaterial som föregår den slutgiltiga utprovningen. Den slutgiltiga utprovningen ska genomföras med tillräckligt många elever, och elever med tillräcklig spridning, för att kunna göra trovärdiga skattningar av hur uppgifterna kommer att fungera när de används i det färdiga provet. Ett vanligt riktvärde här är minst 300 elever, men provramverket ska beskriva de specifika förhållanden som kan råda och påverka detta antal.

Den slutgiltiga utprovningens design och omfattning ska beskrivas och motiveras i provramverket. Bakgrundsinformation ska inhämtas i samband med utprovningen (kön, tidigare betyg, eventuella särskilda upplysningar från läraren m.m.). Uppgifter som ingår i den slutgiltiga utprovningen ska vara av så god kvalitet att ändringar av annat än rent kosmetisk art inte behöver eller ska göras.

Uppgifter utan kända egenskaper ska inte ingå i nationella prov. Elevers och lärares synpunkter inhämtas i samband med såväl mindre som större utprovningar och används i arbetet med justering och val av uppgifter samt i sammansättning och sekvensering av skarpa prov.

5.2.6 Avsedda egenskaper för enskilda uppgifter och för provet i sin helhet

Avsedda egenskaper för enskilda prov formuleras på basis av överväganden kring validitet, reliabilitet och rättvisa. Egenskaperna kan vara av olika typ men ska innefatta relationen till styrdokument, användares uppfattningar samt psyko-metriska egenskaper.

PROVRAMVERK: AVSEDDA EGENSKAPER FÖR PROVET

I provramverket ska avsedda egenskaper för enskilda uppgifter och för provet i sin helhet beskrivas och motiveras. Proqramverket ska också precisera de metoder som ska användas för att fastställa dessa egenskaper. Viktiga aspekter att beakta är provets och provuppgifternas relation till styrdokumentens beskrivning av ämnets långsiktiga mål, centralt innehåll och kunskapskrav, olika intressenters (särskilda granskares, lärares och elevers) uppfattningar, psyko-metriska egenskaper vad gäller reliabilitet och provresultatens stabilitet över tid samt frågor om rättvisa och *bias*.

Minimikrav: I avsnitt 3:4 av detta systemramverk anges minimikrav för mät-precision i de nationella proven.

5.2.7 Slutlig provsammansättning

Delar av prov och enskilda uppgifter konstrueras i regel i delvis separata processer, baserade på gemensamma principer. De olika delarna och uppgifterna sätts i slutänden samman till ett helt prov. Principerna för denna komposition ska klargöras och motiveras av skäl som har att göra med såväl transparens som replikerbarhet.

PROVRAMVERK: BESKRIVNING AV PROVSAMMANSÄTTNING

I provramverket ska den process och de principer som ligger bakom sammansättningen av utprövat provmaterial till ett skarpt prov beskrivas. Det ska också anges och kommenteras i vad mån förändringar görs eller inte görs mellan den slutgiltiga utprövningen och det skarpa provet och vilka effekter dessa förändringar kan ha. De principer för sammanslagning av delar till helheter som görs ska också beskrivas och kommenteras. I provramverket ska granskarnas roll belysas, liksom de analyser som föregår förslag och beslut. Proqramverket ska även innehålla en beskrivning av principer för det aktuella provets relation till tidigare givna och framtida, planerade prov med avseende på stabilitet över tid, när det gäller såväl innehåll som svårighetsgrad.

6. Bedömning och rapportering

I det här avsnittet av systemramverket beskrivs ramar för hur Skolverket och lärosätena ska ge förutsättningar för en korrekt, likvärdig, relevant och användbar bedömning av elevsvar på nationella prov.

Bedömningsanvisningar och rutiner kring bedömning av elevprestationer är nödvändiga för möjligheten att dra valida slutsatser utifrån provresultaten. Bedömningarna ska fånga väsentliga aspekter av den kunskap som eleverna visar i förhållande till provuppgifterna, men det handlar också i hög grad om att minska slumpmässiga effekter som skulle kunna bero på att bedömare gör olika bedömningar av samma arbete och att bedömare inte är konsekventa. Slumpmässiga variationer påverkas också i hög grad av hur skalor som beskriver resultaten konstrueras och hur resultat på olika delar av ett prov aggregeras. Dessutom ska nationella prov ge stöd för slutsatser i förhållande till de syften som proven har, vilket ställer höga krav på kravgränser och kravgränssättning, rapportering och jämförelser med relevanta grupper.

Riktlinjerna är uppdelade på tre områden: bedömning av elevprestationer (6.1), skalor och kravgränssättning (6.2) samt tolkning av provresultat (6.3).

6.1 Bedömning av elevprestationer

Syftet med detta avsnitt är att säkerställa att Skolverket och lärosätena upprättar, dokumenterar och följer procedurer som ger bästa möjliga förutsättningar för en korrekt och konsekvent bedömning av provtagares svar på de nationella proven.

Här beskrivs de krav som ställs på bedömningsanvisningar och hur dessa kan kvalitetssäkras genom utprövning. Här ges även information och riktlinjer för hur bedömningen ska gå till, hur instruktioner till bedömare ska vara utformade, hur stöd för bedömarträning kan och bör se ut samt hur med- och sambedömning kan stärka bedömaröverensstämmelsen. Vidare ges riktlinjer för hur korrekta såväl som felaktiga exempel på elevprestationer ska användas.

Innehållet består av fem delar:

- principer för bedömning
- övergripande bedömningsanvisningar
- bedömningsanvisningar för enskilda uppgifter och provdelar
- utprövning av bedömningsanvisningar
- modeller för korrekt och konsekvent bedömning.

6.1.1 Principer för bedömning

Övergripande principer för bedömning lägger grunden för hur elevprestationerna ska bedömas och är en förutsättning för stabilitet i provsystemet. Det är angeläget att principerna för bedömning är så likartade som möjligt mellan olika nationella prov, eftersom det underlättar när samma lärare genomför bedömningen i flera ämnen och prov och även underlättar elevernas möjligheter att veta vad som bedöms. Principerna bör också vara desamma från ett prov till ett annat inom varje provdel i ett ämne, för att möjliggöra stabilitet i bedömning och jämförbarhet av resultat.

I bedömningen av elevprestationer på nationella prov ska största möjliga objektivitet eftersträvas, och bedömningsanvisningar till de nationella proven ska utformas med detta kriterium i åtanke.

Bedömningen ska utgå från tydliga bedömningsanvisningar, med en konsekvent tillämpning av principer för bedömning. Bedömningen ska bygga på principen om ”positiv rättning” snarare än ”avdragsrättning”, vilket betyder att fokus ligger på att identifiera tecken på visad kunskap snarare än att utgå från ett tänkt fullständigt och korrekt svar och identifiera brister. Bedömningen ska utgå från uppgiftsspecifika bedömningsanvisningar och bedömnarna ska tillämpa bedömningsanvisningens direktiv även om de själva kan ha andra uppfattningar om kvaliteter i elevprestationerna. Bedömningen ska utgå från den prestation som eleven visat, inte från slutsatser om vad eleven möjligen kan ha menat eller från den kvalitet som eleverna brukar visa. Bedömningen ska alltså fokuseras på vad eleverna visar, inte på vad de inte visar.

Om bedömningen handlar om olika dimensioner, får inte elevprestationens kvalitet i en dimension påverka bedömningen av elevprestationen utifrån andra dimensioner.

Bedömningsanvisningarna ska utvecklas parallellt med provuppgifterna och genomgå samma typ av utvecklingsprocess, med bland annat återkommande granskningar i referensgrupper och utprovning (se 5.2.5).

PROVRAMVERK: ÖVERGRIPANDE PRINCIPER FÖR BEDÖMNING

Proqramverket ska innehålla en ämnesspecifik precisering av de principer som allmänt gäller för bedömningen av elevprestationer i provet som helhet och för eventuella delprov.

6.1.2 Övergripande bedömningsanvisningar

Det är angeläget att den information som ges till lärarna inför bedömningen är klar och tydlig, att den innehåller väsentliga inslag som utgör förutsättningar för bedömningsarbetet och att den är konsekvent mellan prov och över tid. Ramar för hur bedömningsprocessen ska gå till ger grundläggande förutsättningar för en robust bedömning av de kvaliteter som eleverna visar i sina svar eller lösningar på uppgifterna i provet. Ramarna kan också minska variationen bedömningsinformationen, vilket underlättar för lärare att bedöma olika prov.

Varje nationellt prov ska åtföljas av en skriftlig information som beskriver bedömningsprocessen allmänt. Informationen ska omfatta viktiga principer för bedömning, generella bedömningsmodeller, konkreta bedömningsanvisningar till provuppgifter (se 6.1.3) och tydliga instruktioner om hur bedömaröverensstämmelsen kan ökas (se 6.1.5).

PROVRAMVERK: INFORMATION OM PRINCIPER FÖR BEDÖMNING

Proqramverket ska innehålla preciseringar av den skriftliga information som ska ges till bedömande lärare inför bedömningen av elevprestationer. Preciseringsen ska visa på vilka principer som är nödvändiga att läraren tillämpar generellt, vilken bedömningsmodell som tillämpas i provet samt vilket stöd som ska ges för förbättrad bedömaröverensstämmelse i det enskilda provet.

6.1.3 Bedömningsanvisningar för enskilda uppgifter och provdelar

Bedömningsanvisningar för enskilda uppgifter och provdelar är en avgörande förutsättning för att bedömningen ska bli likvärdig och korrekt. Oklara anvisningar liksom en variation i utformningen av anvisningar mellan ämnen och årskurser i olika prov lämnar utrymme för godtycke.

Till varje provuppgift ska finnas en beskrivning av vad som krävs av elevernas svar eller lösning för att de ska erhålla poäng eller andra typer av omdömen som beskriver elevsvarets eller elevlösningens kvalitet. Varje sådan beskrivning ska ha ett tydligt format och vara väl anpassad till den provuppgift eller det delprov som den är avsedd för. Bedömningsanvisningarna ska följaktligen vara specifika för de uppgifter och provdelar som de är kopplade till, men sådana specifika anvisningar kan med fördel utgöra varianter av en generell bedömningsanvisning (som i så fall också behöver beskrivas tydligt).

Största möjliga tydlighet ska eftersträvas i bedömningskriterierna, och samtidigt ska tidsåtgången för bedömare att läsa och sätta sig in i bedömningsanvisningarna beaktas. Anvisningarna ska vara så kortfattade och lättillgängliga som möjligt.

Bedömningsanvisningarna kan vara holistiska eller analytiska, och vilken typ som används i ett visst nationellt prov eller delprov ska tydligt beskrivas och motiveras i provramverket. Motiveringen ska särskilt beakta aspekter av tidsåtgång samt hur inter- och intrabedömarreliabiliteten påverkas av den typ av bedömningsanvisningar som används.

Bedömningsanvisningarna ska identifiera godtagbara elevsvar och elevlösningar, dvs. tydligt ange var gränsen går för att ett svar och lösningar ska räknas som godtagbart. Exempel på svar och lösningar av olika kvalitet ska presenteras om det underlättar bedömningen. De svar och lösningar som presenteras ska då primärt illustrera gränsfall mellan olika nivåer av svar eller typer av svar och lösningar som vid utprovning visat sig svårbedömda, och vara väl kommenterade och tydliga så att läraren enkelt kan sätta sig in i vad exemplen illustrerar.

PROVRAMVERK: BESKRIVNING AV BEDÖMNINGSANVISNINGAR

Proqramverket ska precisera utformningen och användningen av bedömningsanvisningarna utifrån Skolverkets angivna ramar och mallar. Det innebär bland annat att principer för utformning av bedömningsanvisningar för olika uppgiftstyper beskrivs och motiveras i proqramverket, och ställningstaganden i förhållande till bedömningsanvisningarnas omfattning. Valet mellan en holistisk eller analytisk bedömningsanvisning ska motiveras, med särskilt fokus på hur tidsåtgången vid bedömningen av elevsvar kan begränsas och hur valet av bedömningsanvisning kan påverka bedömaröverensstämmelsen.

För analytiska bedömningsanvisningar ska proqramverket visa hur bedömningsanvisningarna förhåller sig till olika aspekter av kvalitet. Till exempel kan bedömningsanvisningarna fokusera process och produkt som två kvalitetsaspekter, och en sådan modell bör i så fall finnas och tillämpas konsekvent i hela provet eller provdelen.

Vidare ska användningen av exempel på elevsvar och elevlösningar (korrekta såväl som felaktiga) klargöras. Det kan till exempel handla om mängden elevsvar och elevlösningar, svaren och lösningarnas autenticitet och vilken typ av kommentarer som ska ges till svaren och lösningarna för att de ska stödja bedömningen.

Proqramverket ska också ange förväntad tidsåtgång för bedömningen av elevsvar och elevlösningar, och beskriva vilka överväganden som gjorts för att minimera tidsåtgången med bibehållen validitet. Den totala tidsåtgången för bedömning av svar och lösningar ska vara rimlig och motiverad.

6.1.4 Utprovning av bedömningsanvisningar

Bedömningsprocessen är helt och hållet decentraliserad och bygger på att de instruktioner som skickas ut i form av bedömningsanvisningar har en sådan kvalitet att de ger lärarna möjligheter att göra bedömningar med god kvalitet. Kvaliteten i bedömningsanvisningarna säkerställs genom ett noggrant granskningsförfarande i referensgrupper, men det behövs även empiriska belägg för att bedömningsanvisningarna ska fungera som avsett.

De bedömningsanvisningar som medföljer de nationella proven ska möjliggöra god bedömaröverensstämmelse. Bedömningsanvisningarna ska prövas ut för kvalitetssäkring. Utprovningen ska säkerställa att bedömningen fokuserar viktiga aspekter i elevsvaren och elevlösningarna och även säkerställa en god nivå på bedömaröverensstämmelse för de uppgifter där eleverna ska producera egna svar och lösningar. Mått på prognosticerad bedömaröverensstämmelse ska presenteras i den färdiga bedömningsanvisningen, tillsammans med riktvärden för den bedömaröverensstämmelse som är möjlig att uppnå i det aktuella provet (se även avsnitt 3.4).

PROVRAMVERK: KVALITETSSÄKRING AV BEDÖMNINGSANVISNINGAR

Proqramverket ska precisera hur utprovningen av bedömningsanvisningar ska gå till och vilka mått på bedömaröverensstämmelse som ska anges i förhållande till provet. Proqramverket ska innehålla riktvärden för bedömaröverensstämmelse som ska vara uppfyllda i utprovningen av bedömningsanvisningarna.

6.1.5 Modeller för korrekt och konsekvent bedömning

I ett decentraliserat bedömningssystem är det nödvändigt att stärka kvaliteten genom utbildning och kollegialt stöd. Olika sådana modeller används på skolor, men det är angeläget med ramar för denna verksamhet så att den görs på rätt sätt och så att möjligheterna med sådana modeller tas till vara.

Skolverket ska tillsammans med lärosätena tillhandahålla stöd till hur bedömningen av elevsvar kan kvalitetssäkras genom bedömarträning samt med- och sambedömning. Stödet ska innefatta hur detta kan ske på enskilda skolor, men också principer som kan tillämpas när skolhuvudmän väljer att göra gemensamma bedömningar av elevsvar från flera klasser och skolor.

PROVRAMVERK: MODELLER FÖR BEDÖMNING

Provrामverket ska precisera hur materialet som ger möjlighet till bedömarträning ska se ut och hur instruktionerna för användning av bedömarträning i olika sammanhang ska utformas. Provrामverket ska även innehålla ett resonemang om möjliga modeller för med- och sambedömning för det enskilda provet, och vad som kan åstadkommas med sådana modeller.

6.2 Skolor och kravgränssättning

Syftet med detta avsnitt är att säkerställa att provresultat som ska kunna jämföras genomgår en process som gör dem jämförbara, att jämförelser mellan elevgrupper blir meningsfulla och att den kravgränssättning som görs använder rationella och tydligt beskrivna procedurer.

Innehållet täcks av tre delar:

- skolor och aggregering
- stabilitet och ekvivalering
- kravgränser och kravgränssättning.

6.2.1 Skolor och aggregering

Här anges ramar för hur skolor ska användas i samband med nationella prov och principer för hur provets olika delar vägs samman till en helhet, till exempel i form av ett provbetyg eller en provpoäng. Skalors utformning och processen att aggregera värden påverkar i hög grad möjligheten till reliabla slutsatser baserat på elevprestationerna. Av reliabilitetskäl ska i normalfallet resultaten på de olika delarna i ett nationellt prov aggregeras på ett kompensatoriskt sätt till ett samlat, summativt resultat. Det betyder att mindre bra prestationer i en del kan uppvägas av mycket bra prestationer i en annan del.

För betygsstödjande prov ska det summativa resultaten även uttryckas i betygstermer, ett provbetyg, med hjälp av kravgränser. För prov i årskurser där betyg inte ges kan det summativa resultatet uttryckas i termer av elevens prestation i förhållande till kravnivån för varje delprov. Om andra modeller för beskrivning av elevprestationer används, till exempel redovisning av hur väl elever presterar på enskilda delprov, ska samma krav på reliabilitet ställas på sådana enskilda delprov och delskalor som på provet som helhet. Krav på att

vissa specifika uppgifter ska vara lösta för ett visst provbetyg leder till avsevärda reliabilitetsproblem och ska därför undvikas.

Aggregeringen av resultat i de nationella proven ska göras enligt en kompensatorisk modell, för att minska de reliabilitetsproblem som uppstår med andra modeller (till exempel enumerativa, se Wiliam, 1995). Det innebär att de principer som finns vid betygssättning om att ett kunskapskrav måste vara uppnått i sin helhet för betygen E, C och A inte gäller för proven. Provbetyget eller resultatet ska ge bästa möjliga representation av elevens kunskaper i ämnet i förhållande till kurs- och ämnesplanen inklusive kunskapskraven. Detta säkerställs bland annat genom allsidigt sammansatta prov enligt den provspecifikation som ska ingå i provramverket.

De nationella proven består av delprov. Skalan för olika delprov ska vara gemensam eller kompatibel för att en kompensatorisk modell ska gälla fullt ut. Med kompatibla skalor avses här skalor som på ett meningsfullt sätt kan adderas till ett samlat resultat. Delprovsresultat bör uttryckas i poängskalor som har relativt många steg och som väl representerar variationer i elevresultat. Poängskalorna på enskilda delprov bör inte översättas till delprovsbetyg eftersom övergången till de få stegen i betygsskalan innebär en förlust av information om elevens prestation på delprovet. Översättningen till den begränsade betygsskalan bör göras så få gånger som möjligt för att minimera felen. Delprovsresultaten ska därför aggregeras till ett totalt resultat på provets alla delar innan det tolkas i termer av betygsskalan.

Viktning av olika delprovsresultat bör undvikas, men utifrån en tydlig motivering i provramverket kan viktningar tillämpas. Till exempel kan en provdel ges större tyngd om den kan anses viktigare än andra för att möjliggöra valida tolkningar av resultatet. Användningen av viktning ska tydligt anges i provramverket, tillsammans med välgrundade argument för den viktningssprincip som tillämpas.

Den regel för aggregering av provresultat som fastställs för varje prov ska vara så enkel och genomskinlig som möjligt, och utformningen av regeln ska motiveras i provramverket.

De nationella provens betygsstöd sker genom att varje elevs resultat på provet i sin helhet ges ett provbetyg i skalan A–F (samma som betygsskalan).

PROVRAMVERK: SKALOR OCH AGGREGERING

Provråmverket ska innehålla preciseringar och motiveringar till den skala eller de skalor som används i provet. Provråmverket ska också innehålla en beskrivning och motivering till hur aggregeringen till ett samlat provresultat går till. Provresultat ska rapporteras med skalor som är lämpliga för provets avsedda syfte och då i första hand betygsskalan. Valet av varje poängskala ska motiveras och eventuella procedurer för skalning ska dokumenteras.

6.2.2 Stabilitet och ekvivalering

Olika syften med prov ställer olika krav på hur rigoröst jämförbarheten med tidigare prov måste säkerställas. För det betygsstödjande syftet, och i ännu högre grad för ett utvärderande syfte, ställs relativt höga krav på att likartade prestationer får samma provbetyg på prov inom samma ämne och årskurs eller gymnasial kurs över tid. Därför är det angeläget att ett ramverk för nationella prov behandlar stabiliteten i det nationella provsystemet.

De skalor som används i nationella prov ska ha högsta möjliga stabilitet och arbetet med proven ska ha hög stabilitet som ett tydligt mål. Med stabilitet avses här att variationen i andelen elever med olika provbetyg är minimal, och en acceptabel nivå för resultatens variation ska fastslås. Det är inte rimligt att provresultaten varierar mer än marginellt år från år, åtminstone inte på kort sikt.

Stabiliteten hos de skalor som används ska kontrolleras när provet (inklusive förutsättningar för provet, till exempel kunskapskrav) eller den population av elever som gör provet ändras mer än marginellt. En skala är stabil om positioner på skalan betyder samma sak. Om skalstegens mening förändras ska åtgärder vidtas som minimerar missuppfattningar. Skalans stabilitet ska kontrolleras regelbundet om lämplig användning av provpoängen eller andra provresultat är beroende av skalans stabilitet.

Resultaten på nationella prov ska i största möjliga utsträckning vara jämförbara över tid från ett prov till ett annat inom samma ämne och för samma population. För att resultaten från nationella prov ska vara jämförbara från ett prov till ett annat krävs lämpliga metoder för att länka resultat från olika prov. Det finns olika sätt att göra det på, och olika typer av ekvivalering är olika lämpliga i olika sammanhang. Metoder som bygger på insatta personers värdering av kravgränser bör finnas i det svenska nationella provsystemet, men det är också nödvändigt att införa robusta statistiska metoder för ekvivalering av prov. Bedömningen av provens jämförbarhet ska göras regelmässigt, och denna bedömning ska kompletteras med statistiska metoder.

PROVRAMVERK: STABILITET OCH EKVIVALERING

Proqramverket ska innehålla en beskrivning av hur jämförbarheten mellan ett nytt prov och tidigare givna prov i samma ämne och årskurs eller gymnasial kurs studeras med hjälp av bedömare, till exempel paneler av lärare. Proqramverket ska också beskriva vilka statistiska angreppssätt som används för att ekvivalera proven, till exempel genom användning av ankaruppgifter. Den population av elever för vilken resultaten på de nationella proven ska vara jämförbara ska beskrivas, tillsammans med en övre gräns för hur mycket andelen elever med olika provbetyg rimligen kan variera mellan de olika versionerna av provet. Vilka data som samlas in för att möjliggöra ekvivalering och länkning ska beskrivas, liksom de antaganden som den använda designen utgår från. Proceduren ska beskrivas på en sådan detaljnivå att kunniga personer ska kunna utvärdera och replikera studierna.

Kravgränserna länkar den övergripande bedömningen av kvaliteten i elevernas prestationer på de nationella proven till den bedömning och betygssättning som de nationella proven ska stödja. Kravgränserna översätter provresultaten i betygstermer och det är angeläget att detta viktiga steg har så god stabilitet som möjligt.

Som tidigare nämnts ska de nationella provens betygsstödjande funktion ske via provbetyg. Varje enskild elevs prestation på provet som helhet ska sammanfattas i ett provbetyg som härleds från den skala eller de skalor som används i provet. Resultat från de nationella proven ska vara kompensatoriska till sin karaktär (se 6.3.1), vilket innebär att kravgränser för provbetyg sätts utifrån endimensionella krav, utan särskilda villkor som kräver att eleverna ska ha svarat på ett speciellt sätt på utpekade uppgifter eller delprov. Resultat på delprov sammanfattas företrädesvis inte i betygstermer, utan det är endast provresultatet som

helhet som översätts till betygsliknande omdömen. Grundmodellen är alltså att kravgränser endast ska sättas på de nationella provresultaten som helhet. Om andra modeller tillämpas ska det motiveras tydligt i provramverket.

PROVRAMVERK: KRAVGRÄNSER, PROVRESULTAT OCH PROVBETYG

Provramverket ska innehålla en beskrivning av hur kravgränserna används i det aktuella provet och hur kopplingen mellan provresultat och provbetyg ser ut.

6.2.3 Kravgränser och kravgränssättning

Eftersom de nationella provens betygsstödjande roll i första hand sker via provbetyg är det viktigt att de kravgränser som definierar provbetygen tas fram på bästa möjliga sätt. Vid utvecklingen av ett prov genomförs en kravgränssättningsstudie som ska resultera i ett förslag till kravgränser på det aktuella provet, baserat på empiri från utprovning, kravgränssättningsprocedurer med externa bedömare och bedömningar av kravgränser från referens- och arbetsgrupper.

De metoder som används ska vara väl förankrade i forskning om kravgränssättning och provramverket ska innehålla välgrundade argument för val av metod. Det är angeläget att kravgränssättningen i hög grad grundar sig på data från utprovningar och data från tidigare givna prov, och en framträdande roll för sådana empiriskt grundade modeller ska eftersträvas.

De olika underlagen i kravgränssättningsstudien ska i möjligaste mån vara oberoende av varandra för att understödja triangulering av de föreslagna kravgränserna. Resultat och underlag från kravgränssättningsstudien sammanställs i en rapport. I provramverket preciseras hur kravgränssättningsstudien genomförs. Skolverket tar därefter ställning till om underlaget till förslag på kravgränser följer de riktlinjer som anges i systemramverk och provramverk. Skolverket fastställer därefter kravgränserna. De lärosäten som ansvarar för utvecklingen av de nationella proven ska ha personal med nödvändig kompetens knuten till sig, för att sammanställa det underlag som krävs. Skolverket ska också ha personal med nödvändig kompetens knuten till sig som kan tolka underlaget och avgöra dess kvalitet.

I den kravgränssättningsprocedur med externa bedömare som genomförs i kravgränssättningsstudien är det viktigt att bedömarna förstår provets syfte och även förstår hur kravgränssättningsproceduren går till. Bedömarna ska ha förutsättningar att göra de bedömningar som krävs. Det innebär till exempel att om Angoffs metod, eller någon variant av denna, tillämpas måste bedömarna ha god kännedom om kunskapskrav för olika betygssteg och aktuell kontakt med elever på olika kunskapsnivåer i det ämne och den kurs eller årskurs som provet gäller.

Kravgränssättningsstudien ska dokumenteras på en sådan nivå att personer med kompetens inom området ska kunna utvärdera och replikera studien. I dokumentationen redovisas logiska och empiriska belägg för att kategoriseringarna av elevprestationer i förhållande till kravgränserna är rimliga och försvarbara (validitetsargument). Dokumentationen ska vidare innehålla information om kriterierna för val av bedömare och information om hur de har utbildats inför kravgränssättningen. Dokumentationen ska beskriva om och

hur bedömarna gavs möjlighet att konferera med varandra när de gjorde sina bedömningar, och hur de individuella bedömningarna kombinerades. I varianter av kravgränssättningsprocesser som använder bedömare kan bedömarna få tillgång till data från utprovningar för att stärka stabiliteten i de framtagna kravgränserna. Om sådana varianter används ska de data som bedömarna får tillgång till beskrivas i dokumentationen. Dokumentationen ska innehålla kompletta beskrivningar av de procedurer som följdes och resultaten. Dokumentation ska också, när det är möjligt, innehålla skattningar av den variation som kan förväntas i kravgränsen om studien upprepades med andra bedömare.

Kravgränserna ska åtföljas av mått på precisionen i de kravgränser som används och provramverken ska innehålla riktlinjer för vilken precision som eftersträvas.

PROVRAMVERK: METOD FÖR KRAVGRÄNSSÄTTNING

Proqramverket ska innehålla en beskrivning av den kravgränssättningsprocedur som tillämpas. Principer och procedurer för urval av deltagare och sammanställning av grupper för kravgränssättning beskrivs och motiveras. Proceduren för kravgränssättning beskrivs ingående och motiveras med teori och erfarenhet. En empiriskt och teoretiskt förankrad beskrivning görs i respektive proqramverk av den precision – eftersträvad och uppnådd – som krävs vid fastställandet av kravgränser.

6.3 Tolkning av provresultat

Syftet med det här avsnittet är att bidra till säkerställandet av att provpoäng, andra provresultat och tolkning av information som tillhandahålls i anslutning till nationella prov är begripliga och meningsfulla i förhållande till de avsedda mottagarna. Syftet med avsnittet är inte att begränsa de specifika sätt som provresultat för individer och grupper rapporteras, men att lyfta fram några krav som måste ställas på informationen.

Innehållet består av två delar:

- tolka och förklara provresultat i betygsstödjande syfte
- tolka och förklara provresultat i utvärderande syfte.

6.3.1 Tolka och förklara provresultat i betygsstödjande syfte

En god användning av resultat från nationella prov bygger på att användare förstår vad resultaten betyder och vad de kan användas till. De ramar för sådan information som anges här kan därför bidra till att resultaten användningen nyanseras.

Användare av provresultaten ska få tillgång till den information de behöver för att förstå vad provpoäng och andra provresultat betyder och vilka begränsningar de har. Informationen ska hjälpa användare av provresultat att undvika feltolkningar av resultat för individer och grupper och varna tilltänkta mottagare av informationen för möjliga och troliga feltolkningar av den skala som används. Olika mottagare kan behöva olika typer av information eller liknande typer av information på olika nivåer av tekniska detaljer.

Den information som krävs för att förstå provresultaten ska publiceras samtidigt som provresultaten. Om provpoäng rapporteras ska tilltänka mottagare informeras om begränsningarna hos dessa poäng när det gäller jämförbarhet mellan prov och mellan olika versioner av samma prov. Elever ska också informeras om sina prestationer i förhållande till kravgränser, om det är möjligt och om information hjälper provtagarna.

Informationen ska innehålla beskrivningar av felkällor och risk för felkategorisering i provbetyg samt värderingar av hur säker resultatbilden är på individ- och gruppnivå.

PROVRAMVERK: TOLKNING AV PROVRESULTAT FÖR BETYGSSTÖDJANDE SYFTE

Proqramverket ska innehålla preciseringar av de relevanta typer av information som kan hjälpa användare att tolka provresultat i betygsstödjande syfte.

6.3.2 Tolka och förklara provresultat i utvärderande syfte

En god användning av resultat från nationella prov bygger på att användare förstår vad resultaten betyder och vad de kan användas till. De ramar för sådan information som anges här kan därför bidra till att resultaten användningen nyanseras.

Användare av provresultaten ska få tillgång till den information de behöver för att förstå vad provpoäng och andra provresultat betyder och vilka begränsningar de har. Informationen ska hjälpa användare av provresultat och allmänheten att undvika feltolkningar av resultat för individer och grupper och varna tilltänkta mottagare av informationen för möjliga och troliga feltolkningar av den skala som används. Olika mottagare kan behöva olika typer av information eller liknande typer av information på olika nivåer av tekniska detaljer.

Eftersom nationella prov har ett legitimt allmänintresse ska Skolverket även ge sådan information som hjälper nyhetsmedia och allmänheten att förstå resultaten. Den information som krävs för att förstå provresultaten ska publiceras samtidigt som provresultaten. Informationen ska innehålla beskrivningar av felkällor och risk för felkategorisering i provbetyg samt värderingar av hur säker resultatbilden är på individ- och gruppnivå. Information som ges i lärosätens årliga publicerade rapporter är också en viktig källa till att förstå och tolka resultaten.

Varje nationellt prov ska åtföljas av en eller flera referensramar som kan hjälpa mottagarna att värdera provresultaten. Referensramen kan inkludera information från normstudier, noggrant utvald och definierad statistik på olika nivåer, forskningsstudier eller granskningar av experter. Karaktären hos de grupper som informationen baseras på ska vara tydligt beskriven. Det ska också framgå när det handlar om noggrant utvalda representativa normgrupper och icke-representativa grupper som självutnämnda grupper eller bekvämlighetsurval.

PROVRAMVERK: TOLKNING AV PROVRESULTAT FÖR UTVÄRDERANDE SYFTE

Proqramverket ska innehålla preciseringar av de relevanta typer av information som kan hjälpa användare att tolka provresultat i utvärderande syfte.

7. Riktlinjer för provens användning och genomförande

I denna del av systemramverket anges sådana ramar för provens användning och genomförande som bildar förutsättningar för såväl utveckling som kvalitets-säkring av proven. Genomförandet av nationella prov regleras genom förordning, föreskrifter och hanteringsinformation. Här beskrivs särskilt sådant som lärosätena ska beakta gällande genomförandet av provet.

7.1 Provtider

Provtidens längd ska vara anpassad till elevgruppen. Minst 95 procent av eleverna ska ha tid att besvara uppgifterna i provet på ett tillfredsställande sätt, om inte andra principer kan motiveras. Utprövningar ska ge stöd för att dessa krav är uppfyllda före provtillfället, och i anslutning till provtillfället ska elevernas tidsanvändning undersökas.

PROVRAMVERK: PROVTIDER

Proqramverket ska innehålla principer och motiveringar för den tid som eleverna har till förfogande för att besvara uppgifterna i provet, med motiveringar. I normalfallet ska inte tiden vara begränsande, och om tid görs till en faktor i provet för att till exempel mäta automatiserade förmågor ska detta motiveras särskilt.

7.2 Anpassning av prov och genomförande

Här beskrivs yttre ramar för vilka elever som ska ha möjlighet att genomföra proven och hur eventuella anpassningar ska se ut. Sådana ramar är viktiga för att alla elever ska erbjudas bästa möjligheter att visa sina kunskaper i det ämne som provet handlar om. Ramarna är också viktiga som begränsning av vilka anpassningar som kan göras utan att validiteten i provresultatets användning riskeras.

7.2.1 Allmänt om anpassning av prov

I utformningen av proven ska lärosätena som utvecklar prov sträva efter att så många elever som möjligt ska kunna genomföra proven. Lärosätena ska i provramverket beskriva utgångspunkter för detta och hur de arbetat i den riktningen.

PROVRAMVERK: UTGÅNGSPUNKTER I KONSTRUKTION FÖR MÖJLIGGÖRANDE TILL GENOMFÖRANDE

Proqramverket ska innehålla en beskrivning av de mått och steg som tas i provkonstruktionen för att så många elever som möjligt ska ges möjlighet att visa sina kunskaper.

7.2.2 Inlästa skriftliga provuppgifter

Elever som har nedsatt syn, lässvårigheter eller av andra skäl anses ha svårt att läsa uppgifterna i provet ska erbjudas möjlighet att få uppgifter upplästa. För att standardisera proven ska i normalfallet endast förinspelade uppläsningar användas. Läraren kan annars i sin läsning omedvetet ge ledtrådar som påverkar resultatet. Det är angeläget att elever har möjlighet att lyssna på uppläsningen flera gånger för att lyssnandet i möjligaste mån ska efterlikna de möjligheter som erbjuds läsande elever.

7.2.3 Utökad provtid och annan anpassning vid provtillfället

De elever som inte bedöms ha rimliga förutsättningar att genomföra provet inom provtiden kan erbjudas förlängd provtid om provets syfte och vad som provas tillåter detta. Även andra anpassningar kan vara aktuella. I provramverket ska riktlinjer för anpassningar anges tillsammans med stöd för om och i så fall hur dessa anpassningar ska beaktas vid tolkningen av provresultatet. Provravverket ska även beskriva hur riktlinjerna för anpassning tas fram, hur anpassningarnas konsekvenser för tolkning av resultaten undersöks, och vilka experter på området som deltar i processen.

PROVRAMVERK: MÖJLIGA ANPASSNINGAR VID GENOMFÖRANDE AV PROVET

Provravverket ska innehålla en beskrivning av vilka anpassningar av provets genomförande som kan göras, och under vilka omständigheter anpassningar kan ske. Provravverket ska också beskriva hur tolkningen av provresultatet kan påverkas av anpassningar, samt hur kvaliteten i beskrivningarna av anpassningar säkerställs genom granskning, expertmedverkan och om möjligt utprövning.

7.2.4 Tolkning av provresultat för elever som inte kan genomföra provets alla delar

Provmaterialiet ska innehålla instruktioner för tolkning av provresultat när elever på grund av någon funktionsnedsättning är förhindrade att delta i alla delprov som hör till ett nationellt prov. Det kan till exempel handla om döva elever som inte kan genomföra en del av provet som handlar om hörförståelse. Provets möjligheter till en betygsstödjande funktion i dessa situationer ska beskrivas i provramverket.

PROVRAMVERK: TOLKNING AV PROVRESULTAT

Provravverket ska innehålla en beskrivning av vilket betygsstöd som provet kan erbjuda då elever inte har möjlighet att delta i alla delprov på grund av funktionsnedsättning, även om alla rimliga anpassningar gjorts.

7.3 Provens genomförande, återrapportering och användning

Här beskrivs vad som krävs vid provtillfället för att möjliggöra en valid tolkning och användning av provresultatet.

7.3.1 Tillåten utrustning

Vid genomförandet av ett nationellt prov ska varje elev ha tillgång till den utrustning som provmaterialet anger. Bedömningsanvisningar och kravgränser är utformade utifrån att eleven har, eller inte har, tillgång till viss utrustning och om så inte är fallet kan tolkningarna av elevresultatet ifrågasättas.

Proqramverket ska ange vilken utrustning som varje elev förväntas ha egen tillgång till och vilken utrustning som ska finnas tillhands i rummet.

PROVRAMVERK: TILLÅTEN UTRUSTNING

Proqramverket ska innehålla en beskrivning av den utrustning som eleven förväntas ha tillgång till vid provtillfället och som även funnits tillhands vid utprovningen.

7.3.2 Information om genomförandet

Lärarinformationen som åtföljer det nationella provet ska innehålla beskrivningar av hur genomförandet av provet ska gå till, t.ex. om hur provtillfället ska inledas och avslutas, vilken typ av svar som kan ges på elevernas frågor och hur den personal som deltar under genomförandet bör agera vid provtillfället. Lärarinformationen ska även innehålla instruktioner för hur personal ska agera vid eventuellt fusk och störningar vid provtillfället.

7.3.3 Återrapportering till Skolverket/SCB och lärosäten

För att möjliggöra den nationella uppföljningen av de nationella proven ska provresultat samlas in. Insamlingarna regleras i Skolverkets föreskrifter. Skolverket ansvarar för att återrapportering till Skolverket och SCB kan ske på ett effektivt sätt. De lärosäten som på Skolverkets uppdrag ansvarar för utvecklingen av proven har även möjlighet att göra kompletterande datainsamlingar för att utvärdera och utveckla provens kvalitet. Lärosätena ska utforma datainsamlingar så att de medför minsta möjliga arbetsinsats från skolornas sida.

PROVRAMVERK: UTVÄRDERING AV PROVETS KVALITET

Proqramverket ska innehålla en beskrivning av de utvärderingar som genomförs för att säkerställa och utveckla provens kvalitet.

7.4 Användning av provresultat

Här beskrivs den användning av resultat som förutsätts samt även riktlinjer för hur annan användning kan eller inte kan bli aktuell, och vad som i så fall är nödvändigt att tänka på.

7.4.1 Användning ska primärt kopplas till provens syfte

Proven är utformade under förutsättning att de används på avsett sätt, och denna avsedda användning ska tydligt framgå i provramverk och provmaterial.

PROVRAMVERK: ANVÄNDNING AV PROVRESULTAT

Provramverket ska innehålla en beskrivning av hur provresultaten är avsedda att användas utifrån fastlagda syften.

7.4.2 Annan användning av provresultaten

I den mån provramverket föreslår ytterligare möjligheter att använda provresultaten ska provramverket och informationen i samband med proven visa på stöd för kvaliteten i sådana användningar. Provramverk och information kan också innehålla beskrivningar av vilken användning av provresultat som bör undvikas.

PROVRAMVERK: MÖJLIGHET TILL YTTERLIGARE ANVÄNDNING AV PROVRESULTAT

Om provresultaten på goda grunder kan användas på annat sätt än för fastlagda, gemensamma syften, ska detta tydligt beskrivas och motiveras i provramverket.

8. Referenser

- AERA, AEA & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. ISBN: 978-0-935302-35-6.
- Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3(3), 265–286.
- Dijkstra, J., Galbraith, R., Hodges, B. D., McAvoy, P. A., McCrorie, P., Southgate, L. J., Schuwirth, L. W. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education*, 12(1), 20.
- Downing, S. M. & Haladyna, T. M. (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ebel, R. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational measurement*. Washington DC: American Council on Education.
- Kane, M.T. (2006). Validation. I Robert L. Brennan (Red.), *Educational Measurement* (Fourth edition, s. 17-64). Westport CT: American Council on Education/Praeger Publishers.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.
- Messick, S. A. (1989). Validity. I Robert L. Linn (Red.), *Educational Measurement* (Third edition, s. 13– 103). New York: American Council on Education/ Macmillan.
- Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education*, 14(2), 149–170.
- Newton, P. E. (2010). The Multiple Purposes of Assessment. I E. Baker & B. McGaw (Red.), *International Encyclopedia of Education* (Third Edition) (s. 392–396). Oxford: Elsevier.
- Skolverket (2015). *Provpöängens tillförlitlighet*. Stockholm: Skolverket.
- Wick, J. W. (1987). *School-based evaluation*. Doordrecht: Kluwer.
- Wiliam, D. (1995). Combination, aggregation and reconciliation: evidential and consequential bases. *Assessment in Education: Principles, Policy and Practice*, 2(1), 53–73.



Skolverket

www.skolverket.se