Systemramverk för betygsstödjande nationella bedömningsstöd

Skolverket

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

ISBN: 978-91-7559-347-0

Skolverket, Stockholm 2020.

Innehåll

1. Inledning

Detta systemramverk syftar till att reglera Skolverkets verksamhet med betygsstödjande nationella bedömningsstöd och verksamheten vid de lärosäten som på Skolverkets uppdrag utvecklar sådana material.

Betygsstödjande nationella bedömningsstöd syftar till att vara ett stöd för läraren inför betygssättningen. Bedömningsstöden har formen av ett prov, med eller utan olika delprov. Betygsstödet för läraren består i det provbetyg som genereras baserat på elevens resultat. Förutom bedömningsstöd med ett betygsstödjande syfte tillhandahåller Skolverket även bedömningsstöd med formativa och diagnostiska syften. Föreliggande systemramverk gäller dock endast för de betygsstödjande bedömningsstöden[1].

Betygsstödjande nationella bedömningsstöd ska stödja en rättvis och likvärdig betygssättning i de ämnen, kurser och årskurser som de är avsedda för, och har i det avseendet samma syfte som flertalet nationella prov. Samtidigt skiljer sig betygsstödjande nationella bedömningsstöd från nationella prov till exempel genom att de bygger på frivillighet, inte förnyas med samma regelbundenhet, och att resultaten inte heller sammanställs och publiceras nationellt.

Skolverket har tidigare fastställt ett systemramverk för verksamheten med nationella prov, och föreliggande systemramverk har motsvarande roll för de betygsstödjande nationella bedömningsstöden. Avsikten med systemramverket är att säkerställa högsta möjliga kvalitet i bedömningsstöden och högsta möjliga trovärdighet i resultatens användning i förhållande till bedömningsstödens syfte. Avsikten är också att uppmärksamma önskade och oönskade konsekvenser av bedömningsstödens utformning och användning som en utgångspunkt för kvalitetssäkring.

Systemramverkets primära målgrupp är Skolverket och de lärosäten som på Skolverkets uppdrag konstruerar bedömningsstöd. Föreliggande systemramverk för betygsstödjande nationella bedömningsstöd är en reviderad version av systemramverket för de nationella proven[2], och bygger på ett förslag till systemramverk som utarbetats av universitetslektor Peter Nyström vid Institutionen för didaktik och pedagogisk profession, Göteborgs universitet.

Systemramverket beskriver de aspekter som ska beaktas för att säkerställa kvaliteten i betygsstödjande nationella bedömningsstöd. Systemramverket preciserar också krav på innehåll i det *provramverk* som ska finnas för varje skolämne, årskurs eller kurs där Skolverket tillhandahåller betygsstödjande nationella bedömningsstöd, och krav på innehåll i den dokumentation som ska

[1] De nationella bedömningsstöd i svenska och matematik med diagnostiska syften som är obligatoriska att använda i årskurs 1 omfattas alltså inte av systemramverket.
[2] Systemramverket för nationella prov utarbetades och skrevs av Gudrun Ericksson, Jan-Eric Gustafsson och Peter Nyström, Göteborgs universitet.

åtfölja varje bedömningsstöd i form av en *utvecklingsrapport* (beskrivs senare i systemramverket).

Till varje skolämne, årskurs eller kurs där Skolverket tillhandahåller betygsstödjande nationella bedömningsstöd ska det alltså finnas ett provramverk som lägger grunden för utformning av de specifika bedömningsstöd som utvecklas. Provramverket ska vara stabilt över tid, och vara utgångspunkt för utveckling av nya bedömningsstöd i det skolämne, den årskurs eller kurs som det avser. I provramverket identifieras och diskuteras hot mot validiteten i bedömningsstödet och hur identifierade validitetshot kan minimeras. I provramverket beskrivs även konstruktet, det vill säga tolkningen av vad ett specifikt bedömningsstöd som utvecklas ska ge information om. Här identifieras också avgränsningar, det vill säga vad bedömningsstödet inte avser att pröva. Vidare beskrivs grundläggande förutsättningar för utveckling av bedömningsstöd. Detta kan till exempel röra sig om medverkan av referenspersoner och utprövningsprinciper.

Varje specifikt bedömningsstöd ska dessutom åtföljas av en utvecklingsrapport som anger hur provramverkets ramar följts vid utvecklingen av det specifika materialet. Där ska det finnas en beskrivning av det specifika bedömningsstödets utformning och egenskaper, dokumentation av processen för framtagning av materialet och övrig relevant information om bedömningsstödet (till exempel reliabilitet). Utvecklingsrapporten ska levereras till Skolverket innan ett specifikt bedömningsstöd publiceras, som underlag för beslut om publicering. Skolverket fattar sedan beslut om publicering av bedömningsstödet.

Systemramverket för de betygsstödjande nationella bedömningsstöden (version 1.0) beslutades den 8 december år 2019. Systemramverket kan komma att revideras, exempelvis utifrån förändringar av provsystemet eller politiska beslut.

2. Validitet

Systemramverket för betygsstödjande nationella bedömningsstöd tar sin utgångspunkt i modern validitetsteori (se till exempel Kane, 2006 och Messick, 1989), vilket innebär fokus på användning av prov och andra former för bedömning i vid bemärkelse. Validitet är det mest centrala begreppet inom det pedagogiska bedömningsfältet. Det handlar ytterst om trovärdigheten i de slutsatser, beslut och handlingar som de betygsstödjande nationella bedömningsstöden och deras resultat ger upphov till, och om konsekvenser av bedömningsstödens existens, utformning och användning.

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

Validitetsbegreppet har använts i samband med kvalitet i bedömningar sedan 100 år. Begreppets innebörd har förändrats och utvecklats under denna period 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:

> 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. (Messick, 1989, s. 13)

Denna definition betonar för det första att validitet avser egenskaper i slutsatser och handlingar baserade på bedömningens resultat. Validitet kan alltså inte begränsas till en egenskap hos ett prov utan avser ytterst en adekvat tolkning och användning av provresultaten. För det andra betonar definitionen 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. Validitetsargumenten och validitetsdiskussionen är därmed aldrig avslutad, utan kan ses som en pågående process. För det tredje blir en konsekvens av definitionen att om information från ett prov används för olika syften måste separata bedömningar göras av validiteten i var och en av de olika

användningarna. För det fjärde framgår det explicit i definitionen att den avser såväl provresultat uttryckta i poängform, som andra former för bedömning och andra sätt att sammanfatta resultat av bedömningar.

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.

Det validitetsbegrepp som systemramverket utgår från omfattar kvalitetsaspekter i samband med pedagogisk bedömning i bred bemärkelse. Det handlar i hög grad om trovärdigheten i tolkningar och användning av resultat, om konsekvenser av bedömningsstödens utformning och användning, och om rättvisa och likvärdighet. Det handlar om att minimera systematiska fel (det vill säga att resultatet påverkas av annat än det som avses), men också slumpmässiga fel (det vill säga tillfälligheter som sänker trovärdigheten). Det senare är vad som avses med reliabilitet, som alltså handlar övergripande om hur faktorer av slumpmässig natur påverkar resultatbilder från bedömningsstöden. Urvalet av uppgifter och vem som bedömer prestationens kvalitet är två källor till slumpmässig variation i provresultat. Alltför få uppgifter, och bristande systematik i valet av uppgifter, samt inkonsekvens hos bedömare och mellan bedömare bidrar till slumpmässig resultatpåverkan och sänkt trovärdighet hos resultaten.

Även rättvisa och likvärdighet kan ses som delar i argumenten för trovärdighet och användbarhet i bedömningsstödens resultat, och därmed som delar i validitetsbegreppet. För betygsstödjande nationella bedömningsstöd är det relevant med krav på rättvisa, eftersom resultaten ska stödja betygssättning. Det är av avgörande vikt att bedömningsstöden inte missgynnar grupper av elever utifrån till exempel kön, social bakgrund eller migrationsbakgrund. Det är också angeläget att bedömningsstöden ger elever med olika typer av funktionsnedsättning möjlighet att visa vad de vet och kan göra inom de områden som prövas. Bedömningsstöden ska också innehålla information om hur resultat kan tolkas och hur genomförandet kan anpassas för elever som förhindras att delta fullt ut på grund av funktionsnedsättning. Proven ska vara rättvisande för alla elever som deltar och inte representera något annat än den kunskap provet avser att mäta. Det finns en lång rad källor till validitetsproblem kopplade till bristande rättvisa, som provgenomförandet, provets innehåll, provuppgifternas och svarssättens utformning, bedömningen av elevprestationer och konstruktion av skalor. Bedömningsstöden måste kunna fungera för de elever 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.

2.2 Validering

Modern validitetsteori betonar validering som en process i första hand, och inte så mycket validitet som en egenskap hos ett bedömningsstöd. Det handlar om att motivera och argumentera för kvaliteten i en viss utformning och användning av

ett bedömningsstöd, och om att visa på hur utformningen minimerar risken för icke avsedd tolkning, användning och konsekvenser.

Det första steget i utveckling av ett bedömningsstöd är att precisera den egenskap eller domän som provet ska fånga. I systemramverket kommer termen konstrukt att användas för att representera det bedömningsstödet avser att mäta. Konstruktet kan till exempel utgöras av hela eller delar av det som beskrivs i en kurs- eller ämnesplan, men preciserat och tolkat så att det kan översättas till provuppgifter och bedömningssituationer. 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. Beskrivningen av konstruktet ska tydligt framgå i provramverket för varje bedömningsstöd.

De mest fundamentala frågorna i valideringen är i vilken utsträckning bedömningsstödet förmår att täcka konstruktet i dess helhet, och i vilken grad det innehåller sådant som inte är relevant att pröva. 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. Underrepresentation av konstruktet behöver inte vara ett allvarligt hot mot tolkningar och slutsatser, under förutsättning att det som inte prövas korrelerar väl med det som ingår i provet. Samtidigt kan underrepresentation ge negativa konsekvenser när det gäller påverkan på vad som anses viktigt i ett ämne och vad som tas upp i undervisningen. Om bedömningsstödet endast omfattar delar av det innehåll som ska bedömas är det angeläget att i valideringen beakta i vilken grad och på vilket sätt underrepresentationen är ett hot mot validiteten.

En vanlig källa till konstrukt-irrelevant varians är överutnyttjande av vissa svarssätt, som exempelvis textproduktion, vilket medför att provet tenderar att mäta svarsfärdigheten snarare än den innehållsliga kompetens som provet är avsett att mäta. Konstruktirrelevant varians kan också uppstå genom orimliga krav på den läsförmåga som krävs för att förstå uppgifter. Långa provtider kan också bidra med validitetshot genom att i alltför hög grad mäta uthållighet och inte det innehåll som täcks av uppgifterna. Om bedömningsstödets krav på till exempel textproduktion, läsförmåga och uthållighet bedöms ingå i det konstrukt som prövas är det inte fråga om någon irrelevant varians. Det är endast när tolkning och användning av bedömningsstöden påverkas av sådant som inte ingår eller som kan anses perifert i konstruktet som det blir fråga om en irrelevant varians. Det är angeläget att i valideringen undersöka potentiella källor till konstrukt-irrelevant varians och se till att de inte utgör allvarliga hot mot validiteten.

Validering innebär primärt att pröva hållbarheten i den information som provet ger. Sådana studier ger ofta värdefull information som både kan ge stöd för föreslagna tolkningar och avsedd användning, och ge grund för att ifrågasätta dem. Det är angeläget att olika informationskällor med olika typer av information används i valideringsstudier eftersom de kan belysa olika validitetsaspekter. Sex typer av information är särskilt intressanta och relevanta i valideringsstudier: innehåll (1), svarsprocesser (2), intern struktur (3), relation till andra variabler (4), konsekvenser (5) och reliabilitet (6).

Innehåll (1) handlar främst om relationen mellan undervisningens syfte, centrala innehåll och kunskapskrav som de anges i styrdokument å ena sidan, och provets innehåll och utformning å den andra. När ett bedömningsstöd används som stöd för betygssättning är det en förutsättning att eleverna har haft möjlighet att lära sig det innehåll som ingår i provet.

Information om elevers **svarsprocesser** (2) är också viktig vid validering eftersom det kan finnas flera vägar fram till ett godtagbart svar på en uppgift. Även om det slutliga svaret motsvarar kraven i en bedömningsanvisning kan vägen dit vara oönskad, till exempel genom alltför stora möjligheter att gissa. Svarsprocesser kan också visa på hur elever förstår uppgifter och vad de uppfattar efterfrågas i uppgifterna. 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. Det är vidare av central betydelse att resultaten för olika grupper av elever inte påverkas av konstrukt-irrelevant varians som till exempel har sin grund i uppgifternas och svarssättens utformning.

Med information om **intern struktur** (3) avses resultat av analyser om hur uppgifter och provdelar hänger ihop och tillsammans bidrar till ett trovärdigt och användbart resultat. Det handlar bland annat om hur väl uppgifterna i ett prov korrelerar med varandra, hur väl de diskriminerar mellan nivåer av kunnande och om dimensionaliteten i förhållandet till konstruktet. Om det konstrukt som prövas i ett bedömningsstöd består av ett antal separat identifierbara dimensioner bör det i valideringen 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.

Valideringsinformation kan fås även genom undersökningar av bedömningsstödets resultat i **relation till andra variabler** (4), till exempel sambandet mellan resultatet på ett visst bedömningsstöd och betyg. 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.

Ytterligare en kategori av information kring validitetsproblem handlar om **konsekvenser** (5) 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. Det kan till exempel handla om oönskade effekter på elevernas syn på ämnet, deras intresse för ämnet och bilden av vad det innebär att vara kunnig i ämnet.

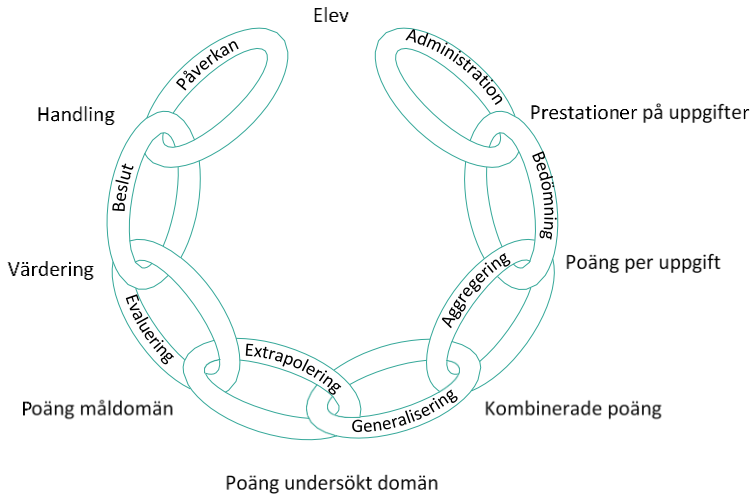
I detta systemramverk ingår även **reliabilitet** (6) som en relevant aspekt i valideringstudier. De metoder som används för att bestämma provets reliabilitet ska vara adekvata för det aktuella provet och för provets avsedda syfte. De variationskällor som påverkar provresultaten ska identifieras, och 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 hel-ler 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.

2.3 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 figur 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 *En modell för pedagogisk bedömning som kan användas vid validering och planering av prov och andra former för bedömning 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 uppmärksammats 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 elevers möte med betygsstödjande nationella bedömningsstöd. Den andra faktorn är provängslan, som snarare är förknippad med hög motivation och kan påverka vissa elevers förmåga att göra sitt bästa i provsituationen. 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 minska genom goda förberedelser inför användningen av bedömningsstödet, 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, till exempel ett prov med fokus på högläsning som 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 (interbedömar- respektive intrabedömarreliabilitet). 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. Uppgifter och procedurer betraktas här som slumpmässiga urval från en mycket större samling uppgifter och procedurer som lika väl hade kunnat ingå i provet. Valideringen handlar här om att bestämma med vilken grad av säkerhet resultat kan generaliseras från de observerade resultaten till den större samlingen uppgifter och procedurer, som här betecknas domän. Resultatens generaliserbarhet avser den korrekthet med vilken de observerade resultaten kan generaliseras till att gälla hela domänen av liknande provuppgifter. 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 eller konstrukt-irrelevant varians. 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 betygsstödjande nationella bedömningsstöd, 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 mätfelen 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. Måldomänen kan likställas med det som i detta systemramverk betecknas konstruktet, dvs. den domän som bedömningsstödet avser att ge information om. 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 flervalsuppgifter 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 **evaluering** 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 resultat från bedömningsstöd inte själv har konstruerat bedömningsstödet 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. Eftersom de betygsstödjande nationella bedömningsstöden är avsedda att stödja en rättvis och likvärdig betygssättning 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 provbetyget. 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 utifrån 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 provramverk och utvecklingsrapport. En ytterligare mer preciserad information finns i efterföljande avsnitt.

2.4 Utformning av validitetsargument

Det samlade validitetsargumentet ska framgå av det provramverk som ska finnas för varje skolämne, kurs eller årskurs som har ett betygsstödjande nationellt bedömningsstöd, tillsammans med den dokumentation i form av utvecklingsrapport som ska åtfölja varje enskilt bedömningsstöd. I provramverket ska bland annat validitetshot identifieras och diskuteras. Där kan också generella principer och åtgärder för att minimera identifierade validitetshot anges. I utvecklingsrapporten ska de konkreta åtgärder som vidtagits i samband med de identifierade validitetshoten beskrivas. Mer ingående beskrivningar av vad som förväntas ingå i provramverk och utvecklingsrapport kommer att presenteras i senare delar av detta systemramverk.

2.5 Samla in och dokumentera validitetsinformationen

De validitetsargument som redovisas i provramverk ska bemöta validitetshot och innehålla relevant validitetsinformation om bedömningsstödet och processen för dess framtagande. Validitetsinformation kan bygga på teoretisk och empirisk grund, och ska lägga grunden för en sammanhängande och övertygande argumentation för användning av bedömningsstödet för det avsedda syftet med den avsedda gruppen av provtagare. 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 elevresultaten. All relevant information ska läggas fram, inklusive sådan som talar mot den föreslagna användningen av provet. Det är inte tillräckligt att validitetsargumentet endast utgörs av en sammanställning av den evidens som råkar vara tillgänglig, oavsett dess relevans. Den samlade evidensen ska enligt resultaten i de genomförda studierna vara tillräcklig för att indikera att bedömningsstödet kan stödja de avsedda tolkningarna av resultatet för att uppfylla avsedda syften.

Utvecklingsrapporter ska granskas i förhållande till system- och provramverk. Skolverket ska leda granskningsprocessen och med utgångspunkt i gjorda granskningar fatta beslut om att bedömningsstöden ska göras tillgängliga för skolor.

Utvecklingsrapporten till varje bedömningsstöd ska beskriva hur reliabilitetsproblem beaktats och motverkats. Där ska även bedömningsstödets reliabilitet anges (se nedan). Den reliabilitetsnivå som krävs för ett visst bedömningsstöd kan endast avgöras genom en professionell bedömning, som beaktar bedömningsstödets syfte och de konsekvenser ett felaktigt beslut kan få.

Nedan anges de typer av reliabilitetsmått som ska preciseras i utvecklingsrapporten för varje betygsstödjande nationellt bedömningsstöd. Kraven avser normalt det totala resultatet, men i den mån stödet för betygssättning endast avser delar av bedömningsstödet gäller kraven för dessa delar.

- Reliabilitet mätt som intern konsistens: Företrädesvis används Cronbachs α som mått. Aktuella mått ska anges i utvecklingsrapporten.
- Medelfel: Beräknas från testinformationsfunktionen (IRT) eller från reliabiliteten (klassisk testteori). Varje bedömningsstöds medelfel ska anges i utvecklingsrapporten tillsammans med kommentarer om provets förmåga att ge pålitlig information för olika prestationsnivåer.
- Klassificeringskorrekthet: Provbetygens överensstämmelse med betyg på kurs eller års-kurs ska undersökas och information om överensstämmelse redovisas i utvecklingsrap-porten. Om inte slutliga betyg finns att tillgå kan preliminära betyg behöva samlas in. Korrektheten i klassificeringen i provbetyg beräknas dels för samtliga betygssteg, dels för betyget F kontra övriga betygssteg. Inga numeriska minimikrav behöver anges, men kommentarer ges om provets

förmåga att stödja korrekt klassificering. Detta kan till exempel studeras genom jämförelser med betygssättning: I vilken utsträckning överens-stämmer antalet elever med olika betyg med antalet elever med olika provbetyg? I vilken utsträckning får elever samma eller olika kurs- respektive provbetyg? Det finns även metoder för beräkning av klassificeringskorrekthet som enbart använder sig av data från ett enskilt prov och som bygger på klassisk testteori och utgår från provet som helhet (se till exempel Livingston & Lewis (1995), eller metoder som bygger på modern testteori och utgår från enskilda uppgifter (se till exempel Lee, 2010). Beräkningar av klassificeringskorrektheten ska anges i utvecklingsrapporten tillsammans med kommentarer om provets förmåga att identifiera korrekt prestationskategori på grupp- och individnivå.

- Reliabilitet mätt som interbedömaröverensstämmelse: Förväntad grad av interbedömaröverensstämmelse ska anges i utvecklingsrapporten.

3. Provutveckling

I detta avsnitt av systemramverket beskrivs hur bedömningsstöden 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), både ämnesbehörighet och gedigen erfarenhet av undervisning inom det ämnesområde som är aktuellt. Kompetensen ska även innefatta erfarenhet av undervisning med elever i relevanta åldrar, med olika socioekonomiska förhållanden, av flerspråkighet samt specialpedagogisk kompetens. Fokus för arbetet är att skapa bedömningsstöd 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 provutveckling[3] förutsätter tydliga specifikationer, iterativ kvalitetskontroll, empirisk information om enskilda uppgifters kvalitet samt utvärdering av prov med bedömningsanvisningar. Detta kräver samverkan mellan 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. I det följande ges kortfattade riktlinjer för utvecklingsarbetet från förarbete fram till dess att uppgifter satts samman till ett färdigt bedömningsstöd. För varje bedömningsstöd ska det finnas ett provramverk och en utvecklingsrapport som

[3] Det finns en rikhaltig litteratur kring provutveckling, såväl av generell som i huvudsak ämnesspecifik art. Litteratur av detta slag är dock alltid till en 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).

följer dessa riktlinjer. Provramverket och utvecklingsrapporten 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 4, Bedömning och rapportering.

3.1 Bedömningsstödets syfte och målgrupp

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

3.2 Bedömningsstödets konstrukt

Bedömningsstödets konstrukt beskrivs med utgångspunkt i gällande kurs- eller ämnesplaner. Även andra utgångspunkter för definitionen av konstruktet bör synliggöras, till exempel läroplan och andra styrdokument, relevant forskning, nationella och internationella exempel samt erfarenheter av lärande, undervisning och bedömning. För de betygsstödjande nationella bedömningsstöden finns kunskapskrav som kategoriseringen i olika betygssteg måste stödja sig på, eftersom betygen ska baseras på samma kunskapskrav.

3.3 Bedömningsstödets format

Elevuppgifterna ska utformas och presenteras så att bedömningens syfte uppfylls på bästa sätt med hänsyn till elevgruppen och bedömningens genomförbarhet. För enskilda uppgifter handlar det bland annat om svarsformat, textmängd, språk och användning av bilder. För bedömningsstödet som helhet handlar det bland annat om struktur (till exempel i form av delprov), omfattning, tidsåtgång och sekvensering (till exempel i form av ökande svårighetsgrad). Bedömningsstödets uppgiftstyper och svarsformer ska alltså utformas och presenteras på basis av överväganden kring validitet, reliabilitet, rättvisa, ändamålsenlighet för sitt syfte och i relation till målgruppen.

3.4 Sammansättning av arbets- och referensgrupper

Utvecklingen av bedömningsstöden förutsätter samverkan mellan 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 slutliga versioner, förslag till betygsgränser (kravgränssättning) och urval av autentiska elevexempel som kommenteras i relation till styrdokumenten. Det är centralt med bredd när det gäller relevanta kompetenser och erfarenheter (se ovan) och aktiva behöriga lärare ska utgöra en betydande del av de grupper som tillsätts. Lärarna ska även

representera olika delar av landet. I provramverket bör principer för de grupper som ska vara representerade i arbetet beskrivas, och i utvecklingsrapporten ska den faktiska sammansättningen av grupper vid framtagning av ett specifikt bedömningsstöd beskrivas och motiveras.

3.5 Steg i processen att utveckla uppgifter till bedömningsstöden: konstruktion, granskning och utprövning

Utvecklingen av ett bedömningsstöd är en iterativ process där justeringar och revideringar görs på grundval av teoretiska överväganden, successiva granskningar, analyser av empiriska data från utprövningar, erfarenheter från autentisk användning av tidigare uppgifter samt en kontinuerlig dialog med användare. Den slutgiltiga utprövningen 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 bedömningsstödet. Ett vanligt riktvärde här är minst 300 elever, men provramverket ska beskriva de specifika förhållanden som kan råda och påverka detta antal.

3.6 Avsedda egenskaper för enskilda elevuppgifter och för provet i sin helhet

Avsedda egenskaper för bedömningsstöden 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 psykometriska egenskaper.

Provramverket om provutveckling

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

Bedömningsstödets konstrukt: I provramverket ska det eller de konstrukt som bedömningsstödet avser att pröva preciseras och motiveras. Provramverket ska även beskriva de processer och kriterier som ska användas för att avgöra bedömningsstödets innehåll. Här anges grad och art av vad i styrdokumenten som prövas respektive inte prövas. Här beskrivs också hur det som prövas eller inte prövas förhåller sig till skrivningar i kurs- och ämnesplaner. Slutligen ska provramverket ange vilken typ av expertis som används vid precisering av provets konstrukt.

Grundläggande principer för bedömningsstödet och dess framtagning: Provramverket ska också beskriva andra grundläggande förutsättningar för

framtagande av bedömnings-stöd. Det kan till exempel handla om principer för sammansättning av de arbets- och referensgrupper som ska involveras i provkonstruktionen, särskilda förutsättningar för be-dömningsstödets distribution och genomförande (till exempel att de ska vara digitala) eller utgångspunkter när det gäller användning av uppgifts- och provformat. Provramverket ska också ange riktvärden för förväntad reliabilitet i form av intern konsistens, klassificeringskorrekthet och bedömaröverensstämmelse.

Bedömningsstödets avsedda egenskaper: I provramverket ska avsedda egenskaper för be-dömningsstödets delar och helhet beskrivas och motiveras. Provramverket ska också precisera de metoder som används 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, psykometriska egenskaper vad gäller reliabilitet och frågor om rättvisa och bias.

Provspecifikation: Presenteras i tabellform.

Utvecklingsrapporten om provutveckling

Bedömningsstödets format och struktur: I utvecklingsrapporten ska det framtagna bedömningsstödets utformning, inklusive uppgiftstyper och svarsformat, beskrivas i förhållande till ramarna som anges i provramverket. Av särskild vikt är beskrivningar av bedömningsstödets möjligheter och eventuella begränsningar när det gäller att täcka det konstrukt som beskrivs i provramverket.

Processen för provutveckling: I utvecklingsrapporten ska processen för konstruktion av det aktuella bedömningsstödet beskrivas. Utvecklingsrapporten ska redovisa vilka grupper och kompetenser som medverkat i utvecklingsprocessen.

Bedömningsstödets egenskaper: I utvecklingsrapporten ska egenskaper hos det framtagna bedömningsstödet beskrivas med utgångspunkt i ramarna som anges i provramverket. Vidare ska utvecklingsrapporten behandla bedömningsstödets möjliga påverkan, såväl avsedd som icke avsedd, på ele-vers möjligheter att visa sina kunskaper och lärares möjligheter att dra adekvata slutsatser av resultaten. Detta innefattar precisioner av olika reliabilitetsmått (se 2.5).

4. 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 i nationella bedömningsstöd.

Bedömningsanvisningar och rutiner kring bedömning av elevprestationer är nödvändiga för möjligheten att dra trovärdiga slutsatser utifrån provresultat. Bedömningarna ska fånga viktiga aspekter av tecknen på kunskap som elever visar i förhållande till uppgifterna, 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 betygsstödjande nationella bedömningsstöd kunna användas för sitt syfte, vilket ställer krav på kravgränser och kravgränssättning.

Riktlinjerna är uppdelade på tre områden: bedömning av elevprestationer (4.1), skalor och kravgränssättning (4.2) samt tolkning av provresultat (4.3).

4.1 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 elevsvar i nationella bedömningsstöd.

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 och hur instruktioner till bedömare ska vara utformade. Vidare ges riktlinjer för hur exempel på såväl föredömliga som bristfälliga autentiska elevsvar ska användas.

Innehållet består av fyra delar:

- principer för bedömning
- övergripande bedömningsanvisningar
- bedömningsanvisningar för enskilda uppgifter och provdelar
- utprövning av bedömningsanvisningar

4.1.1 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 stabila och så transparenta som möjligt. I bedömningen av elevprestationerna ska största möjliga objektivitet eftersträvas, och bedömningsanvisningar till de betygsstödjande nationella bedömningsstöden 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 kunskap snarare än att utgå från ett tänkt fullständigt och korrekt svar och identifiera brister. Bedömningen ska utgå från uppgiftsspecifika

bedömningsanvisningar och bedömarna ska tillämpa bedömningsanvisningens direktiv även om de själva kan ha andra uppfattningar om kvaliteter i elevprestationerna. Bedömningen ska utgå från det eleven visat, inte från slutsatser om vad eleven möjligen kan ha menat eller från vad eleven brukar visa.

Bedömningsanvisningarna ska utvecklas parallellt med provuppgifterna och genomgå samma typ av utvecklingsprocess, med bland annat återkommande granskningar i referensgrupper och utprövning.

4.1.2 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. 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. Ramarna kan också minska variationen i bedömningsinformationen, vilket underlättar för lärare att använda olika bedömningsstöd. Det är också angeläget att Skolverket och de lärosäten som utvecklar bedömningsstöden strävar efter att bedömningsanvisningarna har en gemensam struktur, för att användare lätt ska kunna känna igen sig.

Varje bedömningsstöd ska åtföljas av en skriftlig information som beskriver bedömningsprocessen allmänt. Informationen ska omfatta viktiga principer för bedömning, generella bedömningsmodeller och konkreta bedömningsanvisningar till provuppgifter (se 4.1.3).

4.1.3 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 och inkonsekvent utformning av anvisningar reducerar transparensen och lämnar utrymme för godtycke.

Till varje elevuppgift ska finnas en beskrivning av vad som krävs av elevernas svar eller lösning för att de ska erhålla poäng. Varje sådan beskrivning ska ha ett tydligt format och vara väl anpassad till den elevuppgift 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 bedömningsanvisningarna beaktas. Anvisningarna ska vara så kortfattade och lättillgängliga som möjligt.

Bedömningsanvisningarna kan vara holistiska eller analytiska, och vilken typ (eller vilka typer) som används i ett visst bedömningsstöd 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 utprövning visat sig svårbedömda, och vara väl kommenterade och tydliga så att läraren enkelt kan sätta sig in i vad exemplen illustrerar.

4.1.4 Utprövning 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 fungerar som avsett.

De bedömningsanvisningar som medföljer bedömningsstöden ska möjliggöra god bedömaröverensstämmelse. Bedömningsanvisningarna ska prövas ut för kvalitetssäkring. Utprövningen 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.

4.2 Skalor 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:

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

4.2.1 Skalor och aggregering

Här anges ramar för hur skalor ska användas i de nationella bedömningsstöden och principer för hur eventuella delprov 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 betygsstödjande nationellt bedömningsstöd 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. 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 bedömningsstöd. Provresultatet ska uttryckas i betygstermer som ett provbetyg (A-F), med hjälp av kravgränser.

Om ett betygsstödjande nationellt bedömningsstöd innehåller delprov ska skalorna för de olika delproven vara gemensamma eller kompatibla för att en kompensatorisk modell ska kunna användas. Med kompatibla skalor avses här skalor som på ett meningsfullt sätt kan adderas till ett samlat resultat. Delprovsresultat bör uttryckas i poängskalor 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.

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

4.2.2 Stabilitet och ekvivalering

Betygsstödjande nationella bedömningsstöd har samma syfte som nationella prov när det gäller att bidra till en rättvis och likvärdig betygssättning. De är dock frivilliga för läraren att använda och ingen nationell sammanställning och analys görs av elevresultat. Bedömningsstöden skiljer sig också från nationella prov genom att nya versioner inte alltid tas fram lika frekvent, eller att de inte alls förnyas under längre tid. Detta innebär att även betygsstödjande nationella bedömningsstöd måste efter-sträva högsta möjliga stabilitet, men att systematisk ekvivalering inte är ett generellt krav. Behovet av ekvivalering är mindre påtagligt då bedömningsstöden inte förnyas så ofta. Det är vid bytet av provmaterial som ekvivaleringen blir nödvändig och så länge samma bedömningsstöd används, och inte blir allmänt känt, är resultaten jämförbara. Dessutom innebär frivilligheten stora svårigheter att genomföra och framförallt utvärdera ekvivalering.

Kravgränserna länkar den övergripande bedömningen av kvaliteten i elevernas prestationer på betygsstödjande nationella bedömningsstöd till den betygssättning som de 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 den betygsstödjande funktionen 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. Resultaten ska vara kompensatoriska till sin karaktär (se 4.2.1), 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 provresultaten som helhet. Om andra modeller tillämpas ska det motiveras tydligt.

4.2.3 Kravgränser och kravgränssättning

Vid utvecklingen av ett prov genomförs en kravgränssättningsstudie som ska resultera i ett förslag till kravgränser för provbetyg på det aktuella provet, baserat på empiri från utprövning, 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 utvecklingsrapporten ska innehålla välgrundade argument för val av metod. Det är angeläget att kravgränssättningen i hög grad grundar sig på data från utprövningar (och data från tidigare versioner av bedömningsstöden om sådana finns), 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. I utvecklingsrapporten preciseras hur kravgränssättningsstudien genomförts. Här redovisas även resultat och underlag från kravgränssättningsstudien. Skolverket tar ställning till om underlaget till förslag på kravgränser följer de riktlinjer som anges i systemramverket. Skolverket fastställer därefter kravgränserna. I lärarinformationen som medföljer bedömningsstödet beskrivs kortfattat precisionen i de kravgränser som används.

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. Dokumentationen ska innehålla kompletta beskrivningar av de procedurer som har använts i kravgränssättningsstudien samt resultaten från denna. Dokumentation ska också, när det är möjligt, innehålla skattningar av den variation som kan förväntas i kravgränssättningen om studien upprepas med andra bedömare.

Provramverket ska innehålla riktlinjer för vilken precision som eftersträvas i kravgränserna. I utvecklingsrapporten samt lärarinformationen ska kravgränserna åtföljas av mått på precisionen hos dessa.

4.3 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 betygsstödjande nationella bedömningsstöd 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, utan att lyfta fram några krav som måste ställas på informationen.

En god användning av resultat från betygsstödjande nationella bedömningsstöd 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 resultatanvändningen nyanseras.

Användare av provresultaten ska få tillgång till den information de behöver för att förstå vad provpoäng och andra provresultat betyder och vilka begränsningar dessa 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. Den information som krävs för att förstå provresultaten ska publiceras samtidigt som bedömningsstöden görs tillgängliga.

Provramverket om bedömning och rapportering

Principer för bedömning: Provramverket ska innehålla en ämnesspecifik precisering av de principer som allmänt gäller för bedömningen av elevprestationer i bedömningsstödet.

Beskrivning av bedömningsanvisningar: Provramverket 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 och ställningstaganden i förhållande till bedömningsanvisningarnas omfattning beskrivs och motiveras i provramverket.

Provramverket ska beskriva vilka överväganden som görs för att minimera tidsåtgången för be-dömning av elevsvar med bibehållen validitet. Den totala tidsåtgången för bedömning av svar och lösningar ska vara rimlig och motiverad.

Kvalitetssäkring av bedömningsanvisningar: Provramverket ska precisera hur utprövningen av bedömningsanvisningar ska gå till och ange värden för bedömaröverensstämmelse med utgångspunkt i utprövningar.

Skalor och aggregering: Provramverket ska innehålla precisering och motivering till skalor eller andra former för att representera resultat som ska användas i bedömningsstödet.

Kravgränser och kravgränssättning: Provramverket ska innehålla en beskrivning av den kravgränssättningsprocedur som tillämpas. Principer och procedurer för urval av deltagare och sammansättning av grupper för kravgränssättningar ska beskrivas och motiveras. Proceduren för kravgränssättning ska beskrivas ingående och motiveras med teori och erfarenhet. En empiriskt och teoretiskt förankrad beskrivning ska göras i respektive provramverk av den eftersträvade precision som krävs vid fastställandet av kravgränser.

Utvecklingsrapporten om bedömning och rapportering

Beskrivning av bedömningsanvisningar: Utvecklingsrapporten ska ange de specifika tillämpningar av provramverkets principer för bedömning som används i det framtagna bedömningsstödet. Utvecklingsrapporten ska också beskriva vilka överväganden som gjorts för att minimera tidsåtgången för bedömning av elevsvar med bibehållen validitet. Den totala tidsåtgången för bedömning av svar och lösningar ska vara rimlig och motiverad.

Kvalitetssäkring av bedömningsanvisningar: Utvecklingsrapporten ska precisera hur utprövningen av bedömningsanvisningar har gått till och beskriva förväntad bedömaröverensstämmelse och möjliga sätt att uppnå godtagbar bedömaröverensstämmelse.

Skalor och aggregering: Utvecklingsrapporten ska innehålla en beskrivning och motivering till hur ett provbetyg uppnås i det aktuella bedömningsstödet.

Kravgränser och kravgränssättning: Utvecklingsrapporten ska innehålla dokumentation från krav-gränssättningsstudien och en beskrivning av hur kravgränserna används i det aktuella bedömnings-stödet och hur kopplingen mellan resultat och provbetyg ser ut.

5. Riktlinjer för provens användning och genomförande

I denna del av systemramverket anges sådana ramar för de nationella bedömningsstödens användning och genomförande som bildar förutsättningar för såväl utveckling som kvalitetssäkring av materialen. Här beskrivs särskilt sådant som lärosätena ska beakta gällande genomförandet av provet.

5.1 Provtider

Den tid som eleverna får för att genomföra uppgifterna i bedömningsstödet ska vara anpassad till elevgruppen. Minst 95 procent av eleverna ska ha tid att besvara uppgifterna på ett tillfredsställande sätt, om inte andra principer kan motiveras. Utprövningar ska ge stöd för att dessa krav är uppfyllda före provtillfället.

5.2 Tillgänglighet och anpassning av bedömningsstödens genomförande

I utformningen av bedömningsstöden ska lärosätena som utvecklar dessa sträva efter att så många elever som möjligt ska kunna genomföra dem. Principer för universell design ska följas för att möjliggöra att bedömningsstöden är tillgängliga för så många som möjligt utan särskilda anpassningar. Lärosätena ska i provramverket beskriva utgångspunkter för detta och hur de arbetat i den riktningen.

För elever med behov av anpassning ska anpassningarna utformas så att bedömningsstödet i största möjliga utsträckning prövar det avsedda konstruktet. Personer med specialpedagogisk kompetens och erfarenhet av att arbeta med flerspråkiga elever bör ingå i arbetet med provutvecklingen, såväl generellt som med avseende på anpassningar av skilda slag. Anpassningar kan göras då eleven har behov av sådana. De som har ansvarat för provutvecklingen ska också förse dem som ansvarar för provgenomförande med information om hur anpassningar kan genomföras.

I lärarinformationen 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.

5.2.1 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 prövas tillåter detta. Även andra anpassningar kan vara aktuella. I provramverket ska riktlinjer för anpassningar anges tillsammans med stöd för om och hur dessa anpassningar i så fall ska beaktas vid tolkningen av provresultatet.

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

Lärarinformationen 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 betygsstödjande nationellt bedömningsstöd. Provets

möjligheter att fungera för sitt syfte i dessa situationer ska beskrivas i utvecklingsrapporten.

5.3 Bedömningsstödens genomförande

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

5.3.1 Tillåten utrustning

Vid genomförandet av ett nationellt bedömningsstöd ska varje elev ha tillgång till den utrustning som lärarinformationen 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.

I lärarinformationen ska provkonstruktören ange vilken utrustning som varje elev förväntas ha egen tillgång till och vilken utrustning som ska finnas tillhands vid genomförandet.

5.3.2 Information om genomförandet

Lärarinformationen som åtföljer bedömningsstödet ska innehålla beskrivningar av hur genomförandet av provet ska gå till, till exempel 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.

5.3.3 Uppföljning av resultat

En viktig del av kvalitetsarbetet vid provutveckling är uppföljning av resultat från genomförda prov. Inför varje ny version av ett bedömningsstöd bör information som användningen av tidigare genomförda bedömningsstöd har genererat beaktas.

Provramverket om användning och genomförande

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

Tillgänglighet: Provramverket ska redogöra för hur principer för tillgänglighet efterföljs i provkonstruktionsprocessen.

Anpassningar: Provramverket ska ange riktlinjer för anpassningar tillsammans med stöd för om och i så fall hur dessa anpassningar ska beaktas vid tolkningen av provresultatet.

Utvecklingsrapporten om användning och genomförande

Provtider: Utvecklingsrapporten ska ange tidsåtgång för genomförandet av bedömningsstödet som helhet och dess olika delar.

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

Anpassningar: Utvecklingsrapporten ska innehålla en beskrivning av vilka anpassningar av provets genomförande som kan göras, och under vilka omständigheter anpassningar kan ske. Utvecklings-rapporten ska också beskriva hur tolkningen av provresultatet kan påverkas av anpassningar, samt hur kvaliteten i beskrivningarna av anpassningar har säkerställts genom granskning, expertmedverkan och om möjligt utprövning.

Tolkning av ofullständiga provresultat: Utvecklingsrapporten ska innehålla en beskrivning av vilket betygsstöd som provet kan erbjuda då elever, på grund av funktionsnedsättning, inte har möjlighet att delta i alla provdelar även om alla rimliga anpassningar görs.

Tillåten utrustning: Utvecklingsrapporten ska innehålla en beskrivning av den utrustning som eleven förväntas ha tillgång till vid provtillfället och som även funnits tillhands vid utprövningen.

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.

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.

Skolverket

www.skolverket.se