APPENDIX TILL SKOLVERKETS RAPPORT ”ETT RULLANDE STICKPROVSBASERAT SYSTEM FÖR KUNSKAPSUTVÄRDERING AV GRUNDSKOLANS ÄMNEN”

Ramverk för ett system för uppföljning av kunskapsutvecklingen i grundskolan

Jan-Eric Gustafsson

Underlag utarbetat på uppdrag av Skolverket 2006-05-16

Sammanfattning

Kapitel 1. Inledning och bakgrund

Rapporten redovisar ett uppdrag från Skolverket kring utformning av ett system för rullande utvärdering av elevers kunskapsutveckling. Uppdragets genomförande har styrts av direktiv dels i ett regeringsuppdrag till Skolverket, dels genom uppdragsdirektiv från Skolverket.

Kapitel 2. Erfarenheter från "National Assessment of Educational Progress"

I slutet av 1960-talet infördes i USA under ledning av Ralph Tyler en modell för nationell utvärdering som var innovativ på många sätt. Medan de standardiserade testsystem som var i bruk vid denna tid fokuserade på att ge säkra resultat för enskilda elever var huvudsyftet med "National Assessment of Educational Progress" (NAEP) att ge en bestämning av nivån på kunskaper och färdigheter på nationell nivå. Eftersom elevresultaten i sig inte var intressanta kunde urval göras både av elever och uppgifter (s k matrissampling), vilket innebar att olika elever besvarade delvis olika uppgifter. Härigenom kunde många uppgifter användas utan att den arbetsmässiga belastningen på den enskilda eleven blev alltför stor. Det faktum att urvalen av elever var relativt små gjorde det möjligt att inte endast använda flervalsuppgifter utan även uppgifter med elevproducerade svar som krävde bedömning.

Erfarenheterna av NAEP under 1970-talet visade att programmet kunde ge nationellt och regionalt representativa data om det amerikanska utbildningssystemet. Men samtidigt hade NAEP en undanskymd plats i den utbildningspolitiska diskussionen, och endast i liten grad hade resultaten policymässiga implikationer. En av anledningarna till detta var att NAEPs sätt att rapportera resultaten, där huvudfokus låg på enskilda uppgifter, var mer lämpat för ämnesexperter än för beslutsfattare. Rapporteringen gav inte en sammanfattande överblick över resultaten, och det var svårt att göra jämförelser över tid. Det var också svårt att få överblick över skillnader i resultat för olika undergrupper av elever. Uppläggningen av NAEP gjorde det även svårt att knyta utbildningsresultat till utbildningspolitik och skolpraktik.

Under 1980-talets början kom därför NAEP att få en ny design som baserades på metodologiska landvinningar inom mätläran, vilka gjorde det möjligt att föra över resultat på en och samma mätskala även om eleverna besvarat delvis olika uppgiftsuppsättningar. Härigenom kunde man dra nytta av matrissamplingsteknikens fördelar, samtidigt som nackdelarna förknippade med redovisning på uppgiftsnivå eliminerades. Dessa fördelar hade dock ett pris i form av en betydligt ökad komplexitetsgrad i genomförande av utvärderingarna. För att kunna få ut den information som efterfrågades ur data var det också nödvändigt att utveckla nya, mycket komplexa, analystekniker.

De nyutvecklade metoderna visade sig vara mycket användbara för att skapa överblickbara resultat, och de gjorde det möjligt att undersöka förändringar i nivån av kunskaper och färdigheter över tid. Erfarenheterna visade dock också att den komplicerade tekniken kunde leda till inkorrekta bestämningar av förändringar, varför man valde att som ett parallellt system även fortsättningsvis använda samma uppgifter som användes redan vid 1970-talets början.

Under 1980-talet växte intresset för att följa upp och utvärdera delstatliga utbildningsreformer, liksom att jämföra olika delstaters resultat. Under början av 1990-talet utvidgades NAEP på frivillig bas till delstatsnivån i ett begränsat antal ämnen. I ett senare steg blev utvidgningen till delstatsnivån permanentad. Från år 2001 har NAEP ålagts en ny uppgift, nämligen att vara en kontrollinstans för de utvärderingssystem på elevnivå som delstaterna har ålagts att införa. Detta innebär en kraftig förändring av NAEPs roll i det att det får en tydlig plats inom det ansvarsutkrävande systemet. Beslutet innebär också att den tidigare frivilligheten för delstaterna att delta upphör.

En stor mängd utvärderingar har genomförts av NAEP. Dessa är överlag mycket positiva, även om en återkommande kritik är att NAEP under sin nästan 40-åriga existens vuxit till ett alltför komplext system, och att det ålagts alltför många uppgifter. Andra kritiker har pekat på att utvecklingen av NAEP inneburit att många av Tyler’s ursprungliga idéer om en ”national assessment” har förfuskats genom en insnävning av uppgiftstyper, stark fokusering på vissa ämnesområden, abstrakt rapportering, och en ökad betoning av ansvarsutkrävande funktioner.

Kapitel 3. Erfarenheter från andra nationella utvärderingssystem

Internationellt komparativa studier av elevers kunskaper och färdigheter, som exempelvis de som genomförs av IEA och OECD, har inte i första hand som syfte att vara nationella utvärderingssystem, men för många länder tjänar sådana studier samma funktioner som NAEP gör i USA. De internationella studierna står också inför likartade metodologiska utmaningar som de nationella utvärderingssystemen, nämligen att beskriva nivå och förändring över tid i kunskaper och färdigheter på utbildningssystemnivå.

Tyngdpunkten i IEAs nuvarande verksamhet ligger på genomförande av regelbundet återkommande studier dels av läsfärdighet (PIRLS, ”Progress in International Reading Literacy Study), dels av matematik och naturvetenskap (TIMSS, ”Trends in Mathematics and Science Study”). Syftet är både att jämföra länders resultat och att studera förändring av resultat inom länder över tid. Både TIMSS och PIRLS bygger på de metoder som utvecklats inom NAEP.

Även OECDs undersökning PISA (”Programme for International Student Assessment”), har det dubbla syftet att jämföra länder och studera förändring över tid. I PISA undersöks vart tredje år läsförståelse, matematik och naturvetenskap. Även PISA använder liknande metoder som de som utvecklats inom NAEP.

År 1986 påbörjades det holländska nationella utvärderingsprojektet PPON (”Periodieke Peiling van het Onderwijs Niveau”). Det uttalade motivet bakom införandet av PPON var att ge alla aktörer i det starkt decentraliserade holländska skolsystemet en empirisk grund för sitt agerande.

I sammanfattning kan PPON sägas ha genomgått fyra utvecklingsfaser. Den första fasen inleddes med en undersökning av aritmetik bland 12-åringar och utvidgades sedan att omfatta fler ämnen och upprepade mätningar. En andra fas inleddes 1997, då mätningarna fokuserades på aritmetik och modersmål. År 2003 inträffade en tredje, kortlivad, fas då mätningar i flera ämnen återupptogs eller planerades. Den senaste tiden har dock systemets legitimitet alltmer minskat. En förklaring till detta är att offentlig redovisning av resultat på andra prov som genomförs på individnivå blivit vanliga. En ytterligare förklaring är att det finns stort behov av förklaringar till funna resultat och diskussion kring möjliga handlingsalternativ. PPON är dock inte utformat på ett sådant sätt att det ger underlag för mer fördjupade analyser av orsakerna till de erhållna resultaten.

Även England har vad som kan betraktas som ett nationellt utvärderingssystem, men det är inte urvalsbaserad utan individfokuserat. Basen för det engelska systemet utgörs av en omfattande nationell läroplan med preciserade målbeskrivningar och prov för att mäta uppnående av dessa mål i olika åldrar (7, 11 och 14 år). Ett syfte med proven är att de skall vara diagnostiska och visa på olika elevers starka och svaga sidor. Resultaten sammanställs dock också i syfte att visa resultat på skol- och distriktsnivå. Det engelska systemet har mött mycket kritik och har ifrågasatts av flera olika grupper. I synnerhet gäller detta proven för 7-åringar.

År 1995 startade på Nya Zeeland ett nationellt utvärderingssystem med beteckningen National Education Monitoring Project (NEMP). NEMP är på ett tydligt sätt inspirerat av Tylers ursprungliga idéer om utformningen av NAEP.

Analys och redovisning av resultaten sker på uppgiftsnivå. Efter varje tillfälle publiceras ungefär 2/3 av uppgifterna, medan återstoden återanvänds i nästa utvärderingscykel för att studera förändring över tid. Rapportering görs också för olika undergrupper. Ett utmärkande drag hos NEMP är att projektet i stor utsträckning involverar yrkesverksamma lärare, både för att man vill utnyttja den kompetens lärarna besitter, och för att lärarna måste vara delaktiga för att NEMP skall kunna förbättra undervisningen.

De nationella utvärderingssystemen uppvisar stora skillnader. Vissa är ansvarsutkrävande (England, och i ökande utsträckning NAEP), medan andra är utvecklingsorienterade (NEMP). Detta återspeglar också skillnader i syften och olika prioriteringar vad avser validiteten i beskrivningarna av kunskaper och färdigheter,

Kapitel 4. Beskrivning av kunskaper och färdigheter

Det finns stora metodologiska utmaningar i att beskriva nivån av kunskaper och färdigheter på nationell nivå, och då särskilt att studera förändringar över tid. Dessa avser bland annat validitet, reliabilitet, och generaliserbarhet.

Enligt det vida validitetsbegrepp som framförallt är förknippat med Messick är det viktigt att uppmärksamma även de konsekvenser mätningarna har för individer och grupper, och för vårt sätt att tänka kring de undersökta fenomenen. Ett annat viktigt validitetsproblem är i vilken utsträckning utvärderingssystemet omfattar hela bredden av mål och verksamheter.

Ett ramverk för en utvärdering definierar huvudstrukturen av kunskaper och färdigheter inom ett visst ämnesområde, anger vilka typer av uppgifter som skall användas, hur stor andel av det totala antalet uppgifter en viss uppgiftstyp skall utgöra, osv. Utformningen av ramverket är av fundamental betydelse för validiteten i utvärderingssystemet, varvid fyra aspekter framstår som särskilt viktiga: (1) i vilken utsträckning ramverket har acceptans och legitimitet; (2) i vilken utsträckning ramverket ger signaler som utvecklar skolverksamheten i en önskvärd riktning; (3) i vilken utsträckning ramverket återspeglar undervisningens innehåll och former; och (4) och hur väl ramverket fångar upp olika aspekter av ämnesområdet.

De uppgifter som används skall konstrueras i enlighet med ramverkets specifikationer. Ofta är det dock svårt att konstruera uppgifter och bedömningsanvisningar som förmår att fånga upp djupare aspekter av elevernas förståelse. Andra erfarenheter pekar också på att det är svårt att få en mer fördjupad och nyanserad bild av elevers tänkande och förståelse om man endast förlitar sig på de metoder som är tillgängliga i storskaliga undersökningar.

En nationell utvärdering står och faller med kvaliteten i data. Frågorna kring datakvalitet är i hög grad kopplade till validitetsfrågorna, och i många fall finns ett omvänt samband mellan höga ambitioner vad gäller validitet och kvaliteten i de insamlade data. En anledning till detta är att uppgifter som kräver mer omfattande elevproduktion också tenderar att lämnas obesvarade i större utsträckning än andra uppgifter.

Trendmätning kan åstadkommas både genom att inom ramen för väldefinierade och konstanta ramverk successivt byta ut uppgifter, och genom att använda samma uppgifter i samtliga mätningar. Erfarenheterna visar dock också att resultaten är mycket känsliga för ändringar i de procedurer som används vid datainsamling varför det är angeläget att följa principen ”When measuring change, do not change the measure” oberoende av vilken av de två ansatserna man väljer.

Kapitel 5. Fördjupad analys, tolkning och förklaring

Ett argument för att tolkning och förklaring skall vara ett centralt syfte i ett nationellt utvärderingssystem är att beskrivningarna i sig är ointressanta om undersökningarna inte också bidrar med tolkningar och förklaringar som gör det möjligt att vidta åtgärder som leder till förbättringar.

Svårigheterna att dra slutsatser om orsaker på grundval av tvärsnittsstudier är dock stora. Under förutsättning att adekvata variabler har mätts, kan det dock vara möjligt att med hjälp av statistisk metodik kontrollera för inverkan av störande faktorer. Ett annat problem är att den typ av statistisk analys som är nödvändig för att dra kausala slutsatser från tvärsnittsdata är svår att genomföra, och i synnerhet då komplexa urvals- och mätmodeller används.

Bakgrundsvariabler har en central roll i det fördjupade analysarbetet, och information om sådana inhämtas från elever, lärare och skolledare, och ibland även från föräldrar. Sådana variabler kan används som oberoende variabler i fördjupade analyser, och som föremål för analys och beskrivning. Många av de mest intressanta resultaten från nationella utvärderingar har avsett den bild av skola och undervisning som framträtt genom beskrivningar av olika bakgrundsvariabler. Trots den stora betydelse som tillmäts fördjupad analys och förklaring har dock utveckling av instrument för att mäta bakgrundsvariabler fått liten uppmärksamhet.

Kapitel 6. Rapportering och värdering av resultaten

Under förutsättning att uppgifterna är offentliga ger resultatredovisning på uppgiftsnivå en mycket informationsrik beskrivning, När man använder samma uppgift över flera tillfällen ger denna ansats också god information om förändring över tid. Redovisningar av resultat på de skalor som skapas med hjälp av den moderna mättekniken har stora fördelar ur precisionssynpunkt, och tillåter beskrivning av fördelningar. De siffervärden som presenteras kan dock vara svåra att tolka. För att underlätta tolkningen har olika tekniker utvecklats, t ex så kallade uppgiftskartor, vilka illustrerar vilka uppgifter elever på olika poängnivå kan förväntas klara.

I ökande utsträckning tillhandahålls data, resultat och uppgifter via Internet, där det med hjälp av olika verktyg också är möjligt att skapa egna analyser och rapporter.

Valet av beskrivnings- och rapporteringsmodell är nära kopplat till utvärderingens syften och till vilka typer av uppgifter som används. Det framstår dock som nödvändigt att använda sig av flera olika beskrivningsmodeller samtidigt. Utan sammanfattande, abstraherade, mått är det svårt att göra beskrivningar av förändringar över tid, eller göra fördjupade analyser av exempelvis skillnader i resultat mellan olika grupper av individer. Utan de mer innehållsorienterade uppgiftsnära beskrivningarna är det svårt att mer direkt dra nytta av resultaten i utveckling av praktisk verksamhet.

Kapitel 7. Överväganden och förslag avseende utformning av ett nationellt kunskapsbedömningssystem

I kapitel 7 diskuteras utformningen av ett övergripande ramverk för ett nationellt kunskapsbedömningssystem för den svenska grundskolan.

Av uppdragsbeskrivningen följer att systemets främsta syfte skall vara att ge information om utvecklingen över tid av kunskaper och färdigheter, varvid både nivå och spridning skall uppmärksammas. Ett andra syfte med det nationella kunskapsbedömningssystemet är att ge underlag för analyser av orsaker till förändringar i kunskapsnivåerna.

Grundläggande designfrågor

Enligt direktiven skall samtliga grundskolans ämnen omfattas av den nationella kunskapsbedömningen. Uppdraget bör också tolkas så att inte endast de lätt utvärderingsbara ämnesdelarna skall fokuseras, vilket utgör en betydande utmaning.

Det finns vissa praktiska fördelar med att fokusera kunskapsbedömningen på ett ämne i taget, men ur innehålls- och analyssynpunkter finns fördelar med att snarare strukturera kunskapsbedömningen kring kluster av ämnen som är innehållsligt närliggande. Kunskapsbedömningen bör huvudsakligen genomföras i åk 4 och åk 8, men mycket talar också för att anpassa valet av årskurs efter olika ämnes förutsättningar (t ex förekomst av andra prov, och antal år som ämnet studerats). En grundregel bör vara att kunskaps-bedömningarna skall upprepas vart femte år, men med utrymme för viss flexibilitet.

Ett konkret, om dock tentativt, förslag är att TIMSS regelmässigt används i åk 4 och åk 8 vart fjärde år för att delvis täcka av områdena matematik och naturvetenskap. De ramverk för matematik och naturvetenskap som TIMSS bygger på sammanfaller dock inte fullt ut med de ramverk som skulle konstrueras om en motsvarande svensk undersökning skulle byggas från grunden. Jämförelser bör därför göras med ett svenskt ramverk. För de områden som är underrepresenterade i TIMSS kan då ytterligare uppgifter tillföras utvärderingen.

Utformning av kunskapsbedömningarna

En distinktion görs mellan två olika ansatser till nationell kunskapsbedömning: en som betecknas som uppgiftsorienterad; och en som betecknas som skalorienterad. I den uppgiftsorienterade ansatsen ligger fokus i undersökningsdesign, uppgiftskonstruktion, analys, rapportering och den praktiska användningen av resultaten på en innehållsnära nivå, och den naturliga enheten är den enskilda uppgiften eller en grupp av likartade uppgifter. I den skalorienterade ansatsen ligger fokus på utveckling, analys och tolkning av skalor som representerar nivåer av prestationer inom olika innehållsliga domäner.

Den uppgiftsorienterade ansatsen har begränsningar när syftet är att göra trendmätningar, medan den skalorienterade ansatsen har begränsningar i vilka typer av uppgifter och bedömningssituationer som kan användas. Det är därför inte möjligt att fastslå att systemet skall utformas antingen som ett renodlat skalorienterat eller ett renodlat uppgiftsorienterat system, utan dessa ansatser bör kombineras. Fördelarna med den skalorienterade ansatsen när det gäller trendmätning och beskrivningar av fördelningar för olika undergrupper kan då utnyttjas, och fördelarna med de mer komplexa uppgiftstyperna vad gäller mer ingående och mångfacetterade beskrivningar av elevernas kunskaper och färdigheter kan tas till vara. Dessa två ansatser bör därför kombineras till

en integrerad helhet. Båda ansatserna kan förväntas förekomma inom samtliga ämnesområden, även om tyngdpunkten också kan förväntas vara olika.

Ramverket för ett visst område skall definiera de kunskaper och färdigheter som skall bedömas och skall precisera hur detta skall gå till. Ramverket skall också ange fördelningen och arten av uppgifter som skall användas för konstruktion av skalor, respektive användas på uppgiftsnivå. Ramverket bör utvecklas i en brett sammansatt grupp bestående av lärare, elever, föräldrar, skolledare, och ämnes- och läroplans-experter, där olika synsätt finns representerade och där frågor kring innehåll och uppläggning av kunskapsbedömningen utsätts för en omfattande och mångsidig diskussion.

Erfarenheter från tidigare undersökningar pekar på att det är nödvändigt att med god sannolikhet kunna upptäcka förändringar med effektstorlekar som uppgår till ca 0,10. Detta innebär att det behövs en effektiv stickprovsstorlek om 1570 elever vid var och en av de två mätningarna. På grund av att urvalet är av klustertyp, och att matrissamplings-design används behöver den faktiska stickprovsstorleken dock vara ungefär två till tre gånger så stor.

Ett sätt att öka precisionen i urvalsdesignen är att låta varje elev besvara fler uppgifter. Ett annat sätt är att utnyttja information om bakgrundsvariabler, som föräldrarnas utbildning och nationella bakgrund. Möjligheter finns att för de elever som valts ut att ingå i undersökningen skapa en databas genom att information från olika register läggs samman med information som samlas in med hjälp av frågeformulär och prov. Denna information kan användas för att förbättra precisionen i skattningarna, framförallt genom att den ger information om bortfallets omfattning och art.

Som urvalsmodell föreslås ett flerstegsurval. I första steget väljs skolor, eventuellt inom strata. Därefter görs urval av en, några eller alla klasser inom skolan. Urval av elever inom klasser kan också vara aktuellt, men i de fall en matrissamplingsdesign används är det lämpligt att låta alla elever i klassen ingå, och slumpmässigt distribuera de olika häftena till klassens elever.

Fördjupad analys och förklaring

Även om svårigheterna att nå fram till entydiga slutsatser om kausala relationer är stora, kan ett empiriskt underlag ge bättre förutsättningar för en informerad diskussion om förklaringar till de erhållna resultaten.

Den uppgiftsorienterade ansatsen kan ge möjlighet till kvalitativa analyser av arten av förändring över tid, vilka kan utgöra värdefulla tolkningsunderlag i jakten på förklaringar. Samtidigt ger den skalorienterade ansatsen bättre möjligheter att med statistisk metodik undersöka effekter av determinanter på olika nivåer inom och utom utbildningssystemet.

Även för att förstärka analysmöjligheterna är databaser skapade genom sammanläggning av registerinformation med information insamlad inom den nationella kunskaps-bedömningen av stort intresse. Detta kan exempelvis göra det möjligt att samanalysera resultat på nationella prov med resultaten från den nationella kunskapsbedömningen.

Enkäter till elever, lärare och skolledare är den viktigaste källan till information om olika tänkbara förklaringsfaktorer, och om undervisningens inriktning och uppläggning. Med hjälp av enkäter insamlas också information om viktiga utfall, som exempelvis elevers egna bedömningar av kunskaper och färdigheter, attityder till olika ämnen och ämnes-delar, och motivation att lära.

Former för redovisning av resultat

Den information som ett nationellt kunskapsbedömningssystem skapar är av intresse för ett stort antal grupper som exempelvis allmänheten, elever, lärare, skolledare, beslut-fattare på lokal och nationell nivå, läroplansutvecklare, läroboksförfattare, och lärar-utbildare. De olika informationsbehov som dessa grupper har pekar på att det är nödvändigt att på ett optimalt sätt utnyttja informationen både från den skalorienterade och den uppgiftsorienterade ansatsen.

De siffermässiga resultaten från den skalorienterade ansatsen kan vara abstrakta och svåra att tolka. Ansträngningar bör göras att förse presentationerna av resultatför-ändringar med tolkningsanvisningar som gör siffrorna tolkbara och meningsfulla även för grupper som inte har omfattande erfarenhet av användning av kvantitativa metoder.

Möjligheterna att använda webbaserad rapportering bör utredas. En stor fördel med denna är att användarna själva har möjlighet att med enkla metoder genomföra även komplexa analyser. Dessa analyser kan avse såväl resultat på enskilda uppgifter som resultat på olika skalor.

Ett aktivt analysarbete är en nödvändig förutsättning både för ett framgångsrikt sökande efter förklaringar till de observerade trenderna, och för den fortsatta utvecklingen av instrumenten. Det är därför nödvändigt att den nationella kunskapsbedömningens data används i olika forskningssammanhang, och att resultat presenteras i olika vetenskapliga fora.

Lokal utvärdering

Skolor/skolhuvudmän skall ges möjlighet till att på eget initiativ genomföra lokala utvärderingar vilkas resultat kan relateras till de nationella. Ett sätt att göra detta är att ge stöd för att på lokal nivå använda det uppgiftsmaterial som släpps fritt efter varje utvärdering. Dessa uppgifter kommer att vara omsorgsfullt konstruerade och noga utprövade, och vara försedda med omfattande anvisningar för bedömning av elevsvar. Ett förslag är att Skolverkets Provbank vidareutvecklas genom att uppgifterna och delar av det empiriska materialet läggs in i en databas som är fritt tillgänglig för lokal användning.

Ledningsformer och förankring

Det nationella kunskapsbedömningssystemet måste vara stabilt över lång tid, och det måste ha en bred och djup förankring i alla intressentgrupper. En referensgrupp bör därför skapas med representation av bland andra kommuner, lärarfackliga organisationer, och elev- och föräldraorganisationer.

1. Bakgrund och direktiv

Rapporten redovisar ett uppdrag från Skolverket kring utformning av ett system för rullande utvärdering av elevers kunskapsutveckling. En del av uppdraget har varit att göra en inventering av internationella erfarenheter av nationella utvärderingssystem. En annan del av uppdraget har varit att utveckla ett förslag till ett övergripande ramverk för utformning av ett system för rullande uppföljning av kunskapsutvecklingen i grundskolan.

Utgångspunkten för arbetet utgörs av Skolverkets direktiv för uppdraget, vilka i sin tur har baseras på ett uppdrag från regeringen (Regeringsbeslut 2004-12-22, nr 21) till Skolverket.

Regeringens direktiv till Skolverket

I skrivelsen från regeringen betonas att det nuvarande stickprovsbaserade nationella utvärderingssystemet bör utvecklas till ett system med regelbundet återkommande undersökningar som omfattar samtliga ämnen i såväl den obligatoriska skolan, som de frivilliga skolformerna:

> Skolverket bör prioritera en lösning med rullande undersökningar vid utvecklingen av det föreslagna stickprovsbaserade systemet för kunskapsutvärdering för *samtliga ämnen i grundskolan* och *för ämnen i gymnasieskolan och den gymnasiala vuxenutbildningen*.

Skolverket har dock valt att i det första steget begränsa uppföljningssystemet till att omfatta endast grundskolan.

I skrivelsen från regeringen sägs också:

> Den nationella utvärderingen bör så långt som möjligt utnyttja såväl de erfarenheter som följer av Sveriges deltagande i internationella uppföljningar som resultaten från dessa studier.

Regeringsskrivelsen pekar också på behovet att utveckla metoder för uppföljning av de övergripande målen:

> Regeringen har i olika sammanhang och uppdrag till Skolverket också framhållit behovet av att utveckla metoder för bedömning av övergripande mål som finns i läroplanerna. Ett system för rullande nationella utvärderingar skall eventuellt beakta även denna typ av resultatbedömning.

Skolverket har här valt att i det första steget inte låta uppföljningssystemet omfatta läroplanens övergripande mål.

Skolverkets direktiv

Enligt direktiven från Skolverket skall utvärderingsmodellen ha följande egenskaper:

1. Ge beskrivningar av kunskapsnivåer och utveckling av kunnande över tid inom samtliga ämnen i grundskolan, dels för tidigare år, dels för senare år, enligt ett rullande system där olika ämnen återkommer med viss regelbundenhet.
2. Modellen skall möjliggöra (rimligt) stabila jämförelser av elevers kunskaper och färdigheter även efter att läro- och kursplaner reviderats. Det är ett önskemål att det skall vara möjligt att exempelvis efter större kursplaneförändringar göra förändringar i de utnyttjade instrumenten med bibehållna möjligheter till jämförelser över tid.
3. Vid varje årligt utvärderingstillfälle skall såväl elevers kunskaper och färdigheter i de aktuella ämnena som relevanta bakgrundsfaktorer för elever, lärare och skolor kartläggas.
4. Resultaten från utvärderingen skall i första hand presenteras på nationell systemnivå. Det är dock önskvärt att skolor/skolhuvudmän ges möjlighet till att på eget initiativ genomföra lokala utvärderingar vilkas resultat kan relateras till de nationella.
5. Översiktliga resultat från varje utvärderingstillfälle skall presenteras inom ett relativt kort tidsspann (tolv månader) efter utvärderingen genomförts. Denna resultatpresentation skall innehålla nivåer och trend i kunskaper och färdigheter samt i vissa centrala bakgrundsvariabler. Databasuppbyggnad och rapporteringsrutiner ska dock vara så flexibla att de också medger fortsatta fördjupade och mer förklaringsinriktade analyser av insamlade resultatdata.
6. Datainsamling, datahantering och resultatredovisning skall ske på ett så kostnadseffektivt sätt som möjligt.

I uppdraget från Skolverket preciseras också ett antal punkter och frågeställningar som särskilt skall utredas:

1. Att på ett principiellt plan diskutera hur ramverk för instrumentkonstruktion och mätinstrument skall utformas för att modellens krav skall kunna uppfyllas. Uppdraget innebär dock inte utveckling av ramverk eller av nya eller reviderade mätinstrument som sådana.
2. Möjligheterna att utnyttja jämförande internationella kunskapsstudierna PISA, PIRLS och TIMSS som delkomponenter i ett framtida rullande utvärderingssystem skall diskuteras. Förslaget till ett rullande utvärderingssystem får dock ej utgå från att dessa delkomponenter finns att tillgå.
3. Att diskutera val av mätmetodik, urvalsdesign och urvalsstorlekar såväl generellt som med hänsyn till vilka grupper av ämnen som är under utvärdering, samt diskutera vilken precision resultaten minst bör ha.
4. Belysa för- och nackdelar med olika urvalsdesigner då syftet är att ge möjlighet även till lokal tillämpning, samt ge förslag på alternativa lösningar för de möjliga konflikter som kan uppstå med ett instrument som skall utnyttjas såväl på nationell som på lokal nivå.

5. Att diskutera vilka bakgrundsfaktorer som information bör inhämtas kring vid utvärderingstillfällena, samt föreslå metoder för detta. Enkätmaterial från NU-03 skall vara en utgångspunkt för arbetet i denna punkt.
6. Att diskutera vilka förutsättningar som bör gälla för att översiktliga resultat skall kunna presenteras tolv månader efter utvärderingen genomförts.
7. Att diskutera hur databasuppbyggnad och rapporteringsrutiner skall utformas så att de på ett flexibelt sätt medger fortsatta fördjupade analyser av insamlade resultatdata.
8. Att diskutera lämpliga årskurser, ämneskombinationer och periodicitet för de återkommande kunskapsmätningarna.
9. Att diskutera former och tekniker för utprovning och skalering av nya mätinstrument.
10. Att i samverkan med uppdragsgivaren göra en approximativ kostnadsuppskattning som omfattar bjudning, datainsamling, databasuppbyggnad samt första avrapportering för tre exempelämnen. En minimal nivå med minsta tillräckliga tillförlitlighet och domäntäckning, samt en mer ambitiös nivå skall ingå i förslaget.

Uppläggning och disposition

Uppdraget från Skolverket preciserar också att uppdraget:

> innebär att göra en genomgång av erfarenheter kring utformning av rullande nationella utvärderingssystem, och i detta sammanhang på ett principiellt plan diskutera de möjligheter och svårigheter sådana system erbjuder. Härvid skall särskilt följande problem diskuteras:
> - mätning av kunskaps- och attitydförändring över tid;
> - relationer mellan läroplaner och andra styrdokument å ena sidan och utformningen av utvärderingssystemets instrument å den andra samt
> - för- och nackdelar med de olika tekniska lösningar kring vilka erfarenheter finns.

I kapitel 2 och 3 redovisas därför en översikt av internationella erfarenheter av nationella utvärderingssystem. I första hand fokuseras det amerikanska systemet ”National Assessment of Educational Progress” (kapitel 2) tillsammans med mer begränsad information om projektet ”National Education Monitoring Project” från Nya Zeeland, och PPON från Holland (kapitel 3).

Det finns flera skäl till fokuseringen på NAEP. Ett är att det finns en utomordentligt stor produktion av texter i olika genrer kring detta projekt (bl a resultatrapporter, utvärderingar, utredningar, historieskrivningar, metodrapporter, vetenskapliga artiklar, och forskningsprogram) som sammantaget ger en mycket nyanserad och komplex bild av ett nationellt utvärderingssystems möjligheter och problem. Det faktum att NAEP under sin drygt 40-åriga existens genomgått flera djupgående förändringar gör det också till en rik erfarenhetskälla att ösa ur.

Därefter diskuteras i tre kapitel olika aspekter av utformningen av nationella utvärderingssystem. I kapitel 4 diskuteras olika problem förknippade med beskrivning av kunskaper och färdigheter, med särskilt fokus på bestämning av förändringar över tid. Kapitel 5 fokuserar på problem förknippade med fördjupad analys i syfte att nå fram till tolkningar och förklaringar av de erhållna resultaten, och i kapitel 6 diskuteras former för rapportering och värdering av resultaten. I kapitel 7, slutligen, diskuteras olika aspekter på utformningen av ett system för rullande uppföljning av kunskapsutvecklingen i grundskolan.

2. Erfarenheter från “National Assessment of Educational Progress” (NAEP)

NAEPs historia går tillbaka till 1960-talet. Två primära anledningar kan identifieras till varför man i USA vid denna tid var intresserad av att förbättra informationen om utbildningens resultat på nationell nivå (Vinovskis, 1998). Den ena var den s. k. Sputnikkrisen år 1957, som ledde till en kraftig satsning på utveckling av utbildningssystemets omfattning och kvalitet. Den andra var att såväl Kennedy- som Johnson-administrationerna hade ambitionen att öka den federala nivåns inflytande över det starkt decentraliserade amerikanska utbildningssystemet, där ansvaret för utbildningen i allt väsentligt låg på delstatsnivån.

En kommitté under ledning av Ralph Tyler, USAs mest namnkunnige utvärderingsforskare vid denna tid, fick år 1963 i uppdrag att undersöka olika alternativ för att värdera den amerikanska utbildningens förutsättningar och resultat. I en intervju med Stone (1990) beskrev Tyler uppdraget och dess genomförande på följande sätt:

> In July, 1963, I ran into Francis Keppel at the Cosmos Club in Washington, DC. He was then U.S. Commissioner of Education. He told me that he had looked up the initial legislation in 1868 that established the Office of the Commissioner in the Bureau of Education within the Department of the Interior. The legislation required the Commissioner to report from time to time on the “Progress of Education.” Previous commissioners, he found, reported on *inputs* such as the numbers of teachers, the number of students, and the amount of moneys spent, but not on what progress students were making in learning what schools are expected to teach. He asked me to come up with a feasible plan for obtaining the information needed to report on the progress of education in terms of *outputs*.
>
> I worked out a scheme which I checked with several Fellows at the Century Center – statisticians, educators, psychologists. I sent my proposal plan to Keppel, who checked it with his advisors. They thought it very novel but sound. Keppel then took it to John Gardner, President of the Carnegie Corporation, who checked it with top educational administrators. They said it seemed sound but would be attacked by AASA (American Association of School Administrators) for its likelihood of providing data to support public criticism of the schools.
>
> The Carnegie Corporation decided to support the project and assigned a program associate, Lloyd Morrisett, now President of the John and Mary Markle Foundation, to assist me in getting public approval of the plan, defining educational objectives deemed important by the public, developing appropriate instruments for data collection and trying them out. This took six years and a series of conflicts which were resolved amicably. Finally in June 1969, we turned over the plan, the design, and the instruments to the Education Commission of the States, an official public body that agreed to take responsibility for conducting the National Assessment of Educational Progress. (Intervju med Tyler i Stone, 1990, s 105).

En av de konflikter Tyler hänvisar till i intervjun avsåg frågan om NAEP skulle rapportera resultat enbart på nationell nivå, eller om även rapportering skulle ske för enheter på lägre nivå i skolsystemet, som delstatsnivån eller skoldistriktsnivån. Tyler hade från början tänkt sig att rapporteringen skulle göras både på nationell nivå och på delstatsnivå, men detta mötte starkt motstånd från flera inflytelserika grupper, och då inte endast AASA, utan även från forskare och lärare. Vinovskis (1998) nämner som ett exempel att:

> … the president of the National Council of English Teachers admonished teachers ”to fight tooth and nail to prevent a proposed plan to measure the quality of American education.” (Vinovskis, 1998, s. 6).

Kritiken ledde till att Tyler fick överge planen att rapportera resultat för delstater. Vid sidan av den nationella nivån skulle rapportering ske endast ske för fyra geografiska regioner. Detta, tillsammans med att ansvaret för genomförandet av NAEP lades på Education Commission of the States (ECS), som utgjordes av en sammanslutning av delstater, gjorde att oron för otillbörligt närmande av delstatsnivån lade sig.

En annan kritik som riktades mot det föreslagna utvärderingsprogrammet var att det skulle innebära ett hot mot enskilda elevers integritet och skolarbete. Ett sätt att möta denna kritik var att peka på att de speciella utvärderingsmetoder som utvecklades för NAEP skulle innebära att varje elev endast genomförde en del av uppgifterna, att eleverna skulle vara anonyma, och att endast en mycket begränsad mängd bakgrunds-information om varje elev (kön, ålder, etnicitet, och föräldrarnas utbildningsnivå) skulle samlas in. Som ett svar på den framförda kritiken sade Tyler år 1966:

> This project is encountering some difficulties in getting itself understood. It is being confused with a nation-wide, individual testing program, and several common fears are expressed by those who make this confusion. They note that tests used in a school influence the direction and amount of effort of pupils and teachers. In this way, if national tests do not reflect the local educational objectives, pupils and teachers are deflected from their work. This criticism does not apply to the assessment project because no individual student or teacher can make a showing. No student will take more than a fraction of the exercises. No scores will be obtained on his performance. He will not be assessed at any later time and can gain no desired end, like admission to college or a scholarship.
>
> A second fear is that such an assessment enables the federal government to control the curriculum. This is also a misunderstanding. The objectives to be assessed are those which are accepted by teachers and curriculum specialists as goals toward which they work. They have been reviewed by lay leaders throughout the country so as to include only aims deemed important by publicly-spirited citizens. This project will report on the extent to which children, youth and adults are learning things considered to be important by both professional school people and the informed public.

> A third fear is sometimes raised that this project would stultify the curriculum by not allowing changes over the years in instructional methods and educational goals. It should be made clear that the project will assess what children, youth and adults have learned, not how they have learned it. Hence, the assessment is not dependent upon any particular instructional methods. (Tyler, citerad från Vinovskis, 1998, s 6).

Ett ytterligare sätt att markera NAEPs fristående ställning gentemot läroplaner och det sätt på vilket undervisningen genomförs var att basera populationsdefinitionen på ålder (9, 13, 17 och unga vuxna) snarare än på årskurs.

Samtidigt innebar den alltmer distanserade relationen mellan resultatmätningen i NAEP och skolans resurser och processer minskade möjligheter att använda utvärderings-programmet för analytiska ändamål, och på denna punkt fick NAEP utstå hård kritik både från den lokala skolnivån och från forskare. Katzman och Rosen sade år 1970:

> One gets the impression that CAPE (Committee on Assessing the Progress of Education) in its attention to details of statistical validity, simplicity of administration, and use of a quasi-scientific approach, has lost sight of its major aims. It may seem amazing that such a large undertaking could go so far astray, but this becomes understandable when viewed in the perspective of its growth. Overreacting to early opposition, CAPE has evolved to a point of considerable ambivalence with respect to its original purpose of improving educational decision-making at the local, state, and federal level. It is quite clear that National Assessment will provide little information on the policy issues of the day – the effects of segregation, the effects of decentralization, the effects of resource or curriculum shifts. Nevertheless, lip service is paid to the notion that assessment will improve policy…
>
> The National Assessment Program as it stands today can be criticized on several grounds: 1) measuring questionable educational outcomes with questionable techniques; 2) classifying student subpopulations on largely irrelevant dimensions and/or insufficient detail; 3) neglecting to collect any information on school characteristics which would identify policy-performance relationships. In principle all of these shortcomings can be remedied; however, the institutions for administering the program make such remedy unlikely. We question whether the budget for the program might be shifted to better forms of educational research. (Katzman & Rosen, citerade från Vinovskis, 1998, s 7 – 8).

Det var inte bara den omfattande diskussionen kring programmets syften och inriktning som gjorde att det tog 6 år att utveckla det konkreta förslaget till uppläggning av NAEP. Det visade sig också vara ett omfattande och komplicerat arbete att utveckla de instrument som skulle användas. Vinovskis (1998) noterade att:

> The entire assessment development process took much longer than had been planned, largely due to the unanticipated difficulties in constructing such relatively new and novel instruments in ten subject areas, for four age groups

(including young adults), and reflecting different levels of student competence. (Vinovskis, 1998, s 7).

År 1969 var det dock dags att praktiskt börja använda ett provbatteri för nationella utvärderingsändamål, vilket skulle uppfylla högt ställda psykometriska krav, och som accepterades av dem som skulle använda det.

Den ursprungliga NAEP designen

Den föreslagna utvärderingsmodellen var innovativ på flera sätt. Varje utvärderingscykel skulle fokusera ett eller flera brett definierade innehållsområden, som svarade mot läroplansdefinierade ämnen, men som inte var begränsade till det läroplansbestämda stoffet. Tio ämnesområden skulle behandlas: läsning, skrivning, matematik, naturvetenskap, litteratur, samhällsorienterade ämnen (”social studies” och ”citizenship”), bild, musik, och ”career and occupational development”.

För varje innehållsområde skulle paneler bestående av lekmän bilda konsensusgrupper, som skulle komma fram till lämpliga mål för olika åldersgrupper. Provuppgifter skulle sedan utvecklas, som direkt svarade mot de olika målen. Redan från början fanns sålunda ett starkt krav på innehållsvaliditet i NAEP proven.

På flera sätt var den föreslagna modellen tekniskt innovativ. Användning av flervalsuppgifter prioriterades ned till förmån för uppgifter med elevproducerade svar i syfte att stärka innehållsvaliditeten i proven. Vissa uppgifter bjöds individuellt, medan andra bjöds till mindre grupper av elever. För att nå maximal jämförbarhet skulle alla uppgifter bjudas av särskilt utbildad personal.

Delvis olika uppgifter bjöds till olika elever, enligt en så kallad matrissamplingsmodell. Detta gör det möjligt att få information om en stor mängd uppgifter, utan att den tids- och arbetsmässiga belastningen på den enskilda eleven blir orimligt stor. Tanken var att tidsinsatsen för varje elev som ingick i urvalet skulle uppgå till ca 50 minuter.

Varje utvärderingshäfte skulle innehålla uppgifter från de ämnesområden som var i fokus i denna cykel, med en spridning över lätta, medelsvåra och svåra uppgifter, för att varje elev skulle ha möjlighet att lyckas med åtminstone några uppgifter. Uppgifterna skulle inte endast presenteras i tryckt form, utan också läsas upp via bandspelare, för att även elever med sämre läsförmåga skulle kunna visa sina kunskaper och färdigheter. Detta var också ett sätt att se till att alla elever gavs tillräcklig tid att besvara alla uppgifter.

Som redan nämnts skulle urval göras ur tre åldersdefinierade populationer: 9-, 13- och 17-åringar, liksom även ur en population unga vuxna. Elever i såväl allmänna som privata skolor skulle ingå, liksom de personer i de olika grupperna som inte gick i skola. Resultaten skulle sedan redovisas per ålder och separat för olika demografiska grupper, men aldrig för delstat, skoldistrikt, skola eller individ. I huvudsak skulle andelen korrekta svar per uppgift redovisas. Endast en del av uppgifterna skulle dock offentliggöras,

medan huvuddelen skulle vara hemliga för att göra det möjligt att i framtida utvärderingar undersöka förändringar i kunskaper och färdigheter.

Erfarenheterna av NAEP under 1970-talet visade att programmet kunde ge nationellt och regionalt representativa data om det amerikanska utbildningssystemet utan att detta innebar en implicit läroplan, eller jämförelser mellan delstater, skoldistrikt eller skolor. Men samtidigt pekade erfarenheterna av NAEP under 1970-talet att det hade en undanskymd plats i den utbildningspolitiska diskussionen:

> While the staff of NAEP had tried to be more policy relevant in the 1970s (and some later observers argued that they had succeeded more than had been realized at that time), the general impression among policymakers and educators in the 1970s was that NAEP was not particularly helpful to those in decisionmaking situations. (Vinovskis, 1998, s 9)

En av anledningarna till detta var att NAEPs sätt att rapportera resultaten var mer lämpat för ämnesexperter än för beslutsfattare. Huvudfokus i rapporteringen låg på enskilda uppgifter. Detta var en enkel form av rapportering som inte krävde komplicerade analyser, och där resultatpresentationerna var enkla att förstå, men de var också opraktiska i många sammanhang:

> Subject area experts liked the original item-by-item reports, but such reports were quickly found to be unsatisfactory for communicating with policy-makers and the public. When someone asked ’how are kids doing’, she did not want two hundred answers, one for each item – especially if the same general message was being repeated for most of the items. Beginning around 1974, reports began to provide results in terms of average performance over clusters of related items. (Forsyth, et al., 1996, s 29).

Rapportering av förändring av resultat över tid skedde i termer av kluster av uppgifter som var gemensamma över flera mättillfällen. De skalor i vilka rapporteringen gjordes var dock godtyckliga, och bestämdes av svårighetsgraden av de uppgifter som råkade ingå. Detta gjorde det omöjligt att göra jämförelser över åldrar och över olika uppgifts-kluster. Eftersom 25 % av uppgifterna publicerades efter varje tillfälle blev det också färre och färre uppgifter kvar för jämförelser över tid.

Rapporteringen av resultat på uppgiftsnivå gjorde det också svårt att få överblick över skillnader i resultat för olika undergrupper definierade exempelvis efter etnisk, språkmässig eller socioekonomisk tillhörighet. Som en följd av förändringar i det amerikanska samhället och utbildningspolitiken kom denna typ av information att alltmer efterfrågas. Den ökade etniska mångfalden, och en allt större betoning av vikten av lika tillgång till utbildning för alla samhällsgrupper gjorde att behovet av beskrivning av utvecklingen för olika undergrupper blev allt större. Resurser allokerades också till olika grupper, i form bland annat av specialpedagogiska satsningar och kompensatoriska program. Detta medförde behov av mer detaljerad information om utbildningsresultat, och om effekter av olika utbildningspolitiskt motiverade åtgärder på kunskaper och

färdigheter. Den åldersdefinierade urvalsmodellen gjorde det emellertid svårt att knyta utbildningsresultat till utbildningspolitik och skolpraktik.

Utvecklingen av NAEP under 1980-talet

Dessa, och andra problem som hade att göra med ändrade syften jämfört med den ursprungliga designen, gjorde att det fanns behov av nya metodologiska lösningar. Bock, Mislevy och Woodson (1982) pekade på möjligheten att förbättra NAEP på flera punkter genom att ta i anspråk nyutvecklade mätteoretiska modeller:

> The successes enjoyed to date are due largely to advances in survey methodology, most notably the introduction of multiple-matrix sampling theory. It is now possible to obtain sufficiently precise estimates of attainment over a broad range of skills with minimal expenditures of educational resources.
>
> The remaining challenges of assessment, improving methods of reporting results and monitoring them over time, demonstrate the need for commensurate progress in *measurement* methodology. We believe that IRT models and methods designed for multiple-matrix samples of data offer a solution. The incorporation of the advantages of item response curve measurement models with the economies of matrix sampling designs marks a crucial stage in the maturation of educational assessment (Bock et al., 1982, p 11).

Bock et al., (1982) föreslog sålunda att NAEP skulle gå över från att använda den klassiska mätlärans tekniker, till det som kallas den moderna mätläran, eller Item Response Theory (IRT).

1984 års redesign av NAEP

I samband med anbudstävlan för NAEP 1984 föreslog Messick, Beaton & Lord (1983) vid Educational Testing Service (ETS) en ny design, baserad på den moderna mätläran. Med denna design skulle policyrelevansen i analyserna öka, och problemen att studera förändring av resultaten över tid skulle få bättre lösningar. Den ”andra generationens” NAEP innebar också förändringar i urvalsmodellen genom att både ålders- och årskurs-definierade populationer användes, och i rapporteringsmodellen, genom att man skapade sammanfattande resultatmått för olika innehållsdomäner.

Med tanke på den stora betydelse IRT-tekniken kommit att spela för NAEP och fort-farande spelar i dagens nationella utvärderingssystem finns det anledning att presentera denna teknik något mera utförligt.

”Item Response Theory” (IRT)

IRT är en samlingsbeteckning för en stor klass av modeller, i vilken sannolikheten för ett korrekt svar på en uppgift bestäms som en funktion dels av individens förmåga, dels av olika uppgiftsegenskaper, som uppgiftens svårighetsgrad och diskriminationsförmåga.

Den första IRT-modellen utvecklades av den danske statistikern Georg Rasch (1960). Denna modell, som går under beteckningen Raschmodellen, är den enklaste av IRT-modellerna, i det att den endast inkluderar uppgiftens svårighetsgrad som parameter i modellen. Mer komplexa modeller utvecklades under 1960- och 1970-talen av bland andra Lord (1980). Tvåparametersmodellen inkluderar även en parameter för uppgiftens diskriminationsförmåga, och i treparametersmodellen tas även hänsyn till det faktum att personer med mycket låg förmåga kan besvara en uppgift korrekt genom att gissa det korrekta svaret (Lord, 1980). Det finns även IRT-modeller som hanterar variabler som består av ordnade kategorier, där den så kallade ”partial credit” modellen (Muraki, 1992) ofta används.

Den stora fördelen med IRT-tekniken är att denna gör det möjligt att bestämma uppgifters egenskaper (t ex svårighetsgrad) oberoende av vilken grupp av personer som besvarat uppgifterna, liksom även personers förmåga oberoende av vilka uppgifter de besvarat. Det är bland annat dessa egenskaper som gör att IRT effektivt hanterar data från matrissamplingsdesigner, eftersom olika grupper av personer här besvarar delvis olika uppgifter. Förutsättningen för att det skall vara möjligt att skatta olika uppgifters egenskaper på en och samma skala i en matrissamplingsdesign är att det finns direkta eller indirekta länkar mellan olika uppgifter i form av gemensamma grupper av uppgifter.

Även inom den klassiska mätläran är det möjligt att använda sig av matrissamplings-teknik (Lord, 1962), och det var på den klassiska mätläran som första generationens NAEP byggde. En begränsning är dock att endast medelvärden går att skatta med denna teknik, medan fördelningar av resultat inte är åtkomliga. Som redan nämnts bjuder det också speciella problem att studera förändring över tid med den klassiska mätläran i de fall man inte kan använda sig av samma uppgifter.

Enligt den klassiska mätlärans principer är det en fördel om varje deltagande elev tar så få uppgifter som möjligt, medan den moderna mätläran som redan nämnts förutsätter att det finns ett visst överlapp mellan de uppgifter som bjuds till olika elever. Nya tekniker för konstruktion av häften i matrissamplingsdesignen fick därför utvecklas.

”Balanced Incomplete Block Design”

Bland annat föreslog Messick et al. (1983) en design som kallades ”Balanced Incomplete Block” (BIB) design och det finns anledning att beskriva denna något mer utförligt. På grundval av det totala antalet uppgifter som skall ingå i utvärderingen och det antal uppgifter som varje elev kan beräknas klara inom den tillgängliga tidsramen delas uppgifterna upp i ett antal överlappande enheter. Antag, exempelvis, att det totala antalet uppgifter skulle kräva 175 minuter, och att 75 minuter finns att tillgå för varje elev. Uppgifterna kan då delas upp 7 stycken enheter (”block”, betecknade Block A, Block B, …, Block G), och varje elev får genomföra tre sådana block. Enligt BIB designen sätts de tre blocken samman i sju stycken häften (Häfte 1, Häfte 2, …, Häfte 7) på det sätt som beskrivs i Tabell 1.

Tabell 1. Exempel på en BIB design

Häfte	Position 1	Position 2	Position 3
Häfte 1	Block A	Block B	Block D
Häfte 2	Block B	Block C	Block E
Häfte 3	Block C	Block D	Block F
Häfte 4	Block D	Block E	Block G
Häfte 5	Block E	Block F	Block A
Häfte 6	Block F	Block G	Block B
Häfte 7	Block G	Block A	Block C

Enligt denna design förekommer varje block tillsammans med varje annat block, vilket gör det möjligt att med IRT-teknik bestämma den relativa svårighetsgraden för alla uppgifterna på samma skala. Vidare förekommer varje block i var och en av de tre positionerna, vilket är viktigt i den mån det förekommer ordningseffekter orsakade exempelvis av inlärningseffekter från ett block till ett annat, motivationsförändringar, eller trötthet.

Då varje elev gör få block och det finns många block är det dock svårt att både kombinera varje block med alla andra block och att balansera positionseffekter. Antag exempelvis att det finns totalt 20 block, och att varje häfte kan omfatta två block. En BIB-design skulle då kräva inte mindre än 200 häften, vilket skulle vara praktiskt ogenomförbart. En lösning är då att använda så kallad partiell BIB-design, i vilken varje block matchas med endast en delmängd av blocken. Under förutsättning att det finns antingen direkta eller indirekta länkar mellan alla block är det fortfarande möjligt att göra IRT-skattningar av alla uppgifters svårighetsgrad på samma skala. Antaganden måste dock göras om att ordningseffekter inte förekommer.

”Marginal Maximum Likelihood” skattning

Enligt den plan som föreslagits i anbudet skulle ETS använda tekniker som vid denna tid var relativt väl etablerade för att först bestämma uppgiftsparametrarna för alla de uppgifter som ingick i matrissamplingsdesignen, och sedan skulle förmågeparametrar bestämmas för alla de elever som ingick i utvärderingen med utgångspunkt dels i de avgivna svaren, dels i de skattade parametrarna för uppgifterna. Slutligen skulle de skattade elevvärdena sammanställas och analyseras.

När ETS började bearbeta de data som samlats in i NAEP 1984 visade det sig emellertid att de föreslagna procedurerna inte fungerade. Ett problem var att det inte var möjligt att skatta några förmågevärden för de elever som inte hade något rätt, eller som hade högsta möjliga poäng på de uppgifter de fått, vilket var ca 15 % av de deltagande eleverna. Det andra problemet var att de förmågevärden som man kunde skatta var så influerade av mätfel att de var oanvändbara som grund för analys och resultatpresentation. Anledningen till detta var att varje elev enligt den använda matrissamplingsdesignen besvarade så få uppgifter att osäkerheten i förmågeskattningarna var stor i förhållande till variationen mellan olika elever.

I detta krisläge utvecklade ETS nya analysmetoder, med så kallad marginal maximum likelihood (MML) teknik i centrum. Grundidén med MML-tekniken är att man undviker att skatta individuella förmågeparametrar och istället skattar parametrar (t ex medelvärde, standardavikelse) för populationsfördelningarna. Härigenom löser man både problemet att individuella förmågeparametrar inte alltid går att skatta, och att de är felbemängda. Detta innebar en betydande utvecklingsinsats (Mislevy 1987, 1991; Mislevy & Sheehan, 1987, 1989). Bland annat förde man in kolateralinformation i form av bakgrundsvariabler för att öka precisionen i skattningarna, och man introducerade en teknik med ”plausibla värden” för att göra det möjligt att genomföra MML-analyser i sekundäranalyser, utan att man skulle behöva använda hela den exceptionellt komplicerade analysapparaten.

Man löste sålunda problemet, men fick betala ett pris i form av ett mycket komplext, tidsödande och datakrävande analysmaskineri. Detta analysmaskineri har dock visat sig vara svårt att ersätta, och det, eller något moderniserade varianter, används fortfarande i NAEP och har också fått spridning till andra sammanhang, och då framförallt de internationella studierna av kunskaper och färdigheter (t ex PISA och TIMSS).

MML-skattning sker genom en flerstegsprocess. I det första steget skattas uppgifts-parametrarna, vilket kan ske exempelvis med de program som nämnts ovan. I det andra steget genomförs sedan MML-skattningen, och i detta steg betraktas uppgiftspara-metrarna från det första steget som fixa och givna. Det andra steget består av två delsteg. I det första steget genomförs en skattning av regressionen av den latenta förmåge-variabeln på en uppsättning oberoende variabler (t ex skol-, lärar-, och elev-variabler), och det andra delsteget beräknas skattningar av de olika populationsparametrarna (t ex medelvärden för olika subgrupper, standardavvikelser, och percentiler).

Förutsättningen för att denna metod skall ge korrekta resultat är att de klassifikations-variabler enligt vilka resultaten skall rapporteras finns med som oberoende variabler (dummyvariabler) i det första delsteget. Detta innebär att man i allmänhet måste inkludera ett mycket stort antal variabler, vilka ofta också är högt interkorrelerade. I sin tur innebär detta att beräkningarna blir komplexa och tidsödande. Ett sätt att komma förbi dessa problem är att reducera informationen om de oberoende variablerna till okorre-lerade principalkomponenter, vilka sedan används som oberoende variabler i det första delsteget.

Beräkningen av de olika populationsmåtten i det andra delsteget är baserade på en uppsättning så kallade *plausibla värden*, vilka skall betraktas som slumpmässiga värden dragna från individens sannolikhetsfördelning av förmågevärden. I allmänhet beräknas 5 plausibla värden för varje individ. För att dessa i sin tur skall ge skattningar av populationsparametrarna måste de efterbehandlas med speciell programvara.

Det visade sig att det var möjligt att med den nyutvecklade MML-tekniken få fram intressanta resultat inte endast ur NAEP 1984 utan även ur de tidigare genomförda undersökningarna:

> … the IRT model and distributional estimation methods were applied to map the historical stream of pre-1984 reading data into the 0 – 500 scale. Stable and credible results were attained, which echoed, amplified, and made more comprehensible the cross-year and cross-age results that had been reported in the past in terms of average percents-correct. (Forsyth, et al., 1996, s 32-33).

Tanken var att man med 1984 års undersökning skulle skapa en brygga mellan de mätningar som genomförts före 1984 och de kommande mätningarna, och att man därigenom skulle kunna få en fortsättning på den trendlinje som startade med 1970 års undersökning. I 1984 års undersökning samlade man därför in data både enligt den nya BIB-designen och med de procedurer som använts tidigare. Ambitionen att skapa en fortsatt trendlinje kom dock på skam:

> No new trend line ever materialized. Changes in frameworks, item specifications, definitions, and administration conditions were introduced every one or two assessment cycles so that a consistently defined metric could never be established for solid comparisons over time for more than two assessments. (Forsyth et al., 1996, s 33).

Inom NAEP har man därför bevarat de ursprungliga uppgifterna och procedurerna inom ramen för parallella undersökningar av utvecklingen över längre tid och anledningen är att man stött på problem när det gäller stabiliteten i de BIB-baserade mätningarna. Det har dock inte varit utan sin dramatik att komma fram till insikten om att det finns bestämda gränser för vilka förändringar som kan göras av innehåll och procedurer i utvärderingarna, och ändå ha kvar möjligheterna till jämförelser över tid.

1986 års anomali

En särskilt dramatisk händelse var det som kallas ”the 1986 reading anomaly”. Resultaten från läsundersökningen år 1986 visade, särskilt för 17-åringar, på en nedgång i resultaten som var större än någon annan förändring under NAEPs historia. Även om förändringen inte var stor i absoluta termer, bedömdes resultaten ändå som orimliga mot bakgrund av att förändringar på populationsnivå alltid är små över korta tidsperioder. Ingående analyser visade att den siffermässiga nedgången var att hänföra till en serie förändringar i sammansättningen av block inom häften, administrationsprocedurer, och uppgiftskontext (Beaton & Zwick, 1990). Var och en för sig var förändringarna välmotiverade, och syftade till att förbättra designen, men den sammantagna effekten blev att jämförelse-möjligheterna över tid förstördes. Beaton och Zwick formulerade på grundval av dessa erfarenheter principen: ”When measuring change, do not change the measure”.
En konsekvens av 1986 års anomali var att planerna på att fasa ut de procedurer som använts före år 1984 sköts på framtiden. Istället etablerades ”long term trend” (LTT) studier som en fristående komponent vid sidan av huvud NAEP. Det finns en lång rad skillnaderna mellan de två delarna av NAEP.

I LTT NAEP används bandinspelade instruktioner för eleverna, dels i syfte att minska inflytandet av läsförståelse, dels i syfte att kontrollera den mängd tid som allokeras till

varje uppgift. De BIB-spiraler som började användas i huvud NAEP från år 1984 medförde att elever i samma klass hade olika häften, med påföljd att man gick över till att använda skriftliga instruktioner.

En annan skillnad är att LTT NAEP använder sig av åldersdefinierade urval (9, 13, och 17 år), medan huvud NAEP använder årskursdefinierade urval (4, 8 och 12). Inom LTT NAEP gör man ingen översampling av olika undergrupper, vilket sker inom huvud NAEP. En konsekvens av detta är att de trendlinjer som skattas för olika etniska undergrupper inom LTT NAEP enligt en studie av Barron och Koretz (1996) är ”insufficiently precise”. En annan skillnad är att de elever som deltar i LTT NAEP gör uppgifter från olika ämnesområden, medan eleverna i huvud NAEP endast gör uppgifter från ett område.

Inom LTT NAEP används exakt samma uppgifter, vilka till stor del utgörs av flervalsuppgifter, över tid:

> The trend assessments are based on the content frameworks that were developed for the 1983-1984 assessments in reading and writing or the 1985-86 assessments in mathematics and science. Since the development of these frameworks, substantial changes have occurred in the objectives that content experts believe teachers should emphasize. The current practice is to make the changes in the main NAEP assessment called for by content experts and supported by the National Assessment Governing Board, but to leave the trend assessment frameworks undisturbed. (Barron & Koretz, 1996, s 215).

Det har sålunda varit en tanke med huvud NAEP att ramverk och uppgifter skall återspegla en aktuell syn på ämnesområdet, vilket gör det nödvändigt att göra förändringar, vilka i sin tur försvårar jämförelser över tid. Samtidigt är det ett uttalat syfte att även huvud NAEP skall ge underlag för studium av förändring över åtminstone kortare tidsperioder. Möjligheter och problem att mäta förändring i de två delarna av NAEP diskuteras vidare i kapitel 4.

NAEP på delstatsnivå

År 1983 publicerades larmrapporten ”A nation at risk” i vilken det hävdades att man identifierat djupgående kvalitetsproblem i det amerikanska skolsystemet. Rapporten ledde bland annat till delstatliga utbildningsreformer, med ty åtföljande behov av att följa upp och utvärdera dessa. Både på nationell nivå och delstatsnivå utvecklades också ett mycket större intresse för att jämföra olika delstaters resultat. Vid mitten av 1980-talet tog en grupp om åtta delstater, däribland Arkansas med guvernör Clinton i spetsen, initiativ till att på försök använda vissa av NAEPs instrument för utvärderingsändamål (se Vinovskis, 1998, s 12).

Intresset för att utvidga NAEP till delstatsnivån var en anledning att en kommission tillsattes år 1986 med uppgift att göra en översyn av NAEP. Med anledning av att

genomförandet av NAEP nu var förlagt till den privata organisationen ETS var en del av uppdraget att föreslå en förändrad ledningsorganisation för NAEP. Kommissionen leddes av guvernör Lamar Alexander och Thomas James, som tidigare varit ordförande för Spencer Foundation, och bestod av 22 namnkunniga politiker och forskare, däribland Hillary Rodham Clinton, Linda Darling-Hammond, Pascal Forgione, Francis Keppel och Michael Kirst.

Alexander-James kommissionen överlämnade sin rapport år 1987 till Department of Education, i vilken man föreslog att NAEP skulle utvidgas till att även omfatta insamling, analys, och presentation av resultat på delstatsnivå (Alexander & James, 1987). Jämfört med de tankar som låg till grund för den ursprungliga NAEP-designen innebar detta en dramatisk politisk kursändring.

Department of Education remitterade Alexander-James rapporten till National Academy of Education, som utsåg en kommitté under ledning av Robert Glaser att kommentera rapporten och dess förslag. Kommittén ställde sig i många avseenden positiv till rapporten, men uttryckte tvivel om i vilken utsträckning NAEP kan bidra med informationsunderlag för förklaringar av resultat och förbättringar av skolan:

> What is less clear in the panel report is how NAEP data will actually link to school improvement efforts. Although NAEP can tell us a great deal about "how our schools are doing," it provides only limited and mostly indirect evidence about the factors contributing to these successes and failures. It is natural to suggest that NAEP data collection be expanded so as to shed more light on these causal linkages. Unfortunately, few such questions are well suited for examination within the current NAEP design … (Vinovskis, 1998, s 15).

Även om inte kommittén direkt avvisade redovisning på delstatsnivå menade man dock att man fäst för stort avseende vid dessa, och pekade på att delstatsresultaten påverkas av en lång rad faktorer. Trots att det är möjligt att på statistisk väg kontrollera för en del av dessa menade man att jämförelser mellan delstater inte kan förväntas ge bidrag till utveckling av skolan. Däremot pekade man på att jämförelser över tid på delstatsnivå kan vara mer informativa.

Efter en hel del ytterligare politisk diskussion fattade senat och kongress beslut år 1988 om att NAEP på försök år 1990 skulle utvidgas att omfatta de delstater om önskade delta, men endast i matematik för åk 8. Försöksverksamheten skulle sedan utvidgas år 1992 till att även omfatta matematik och läsning i åk 4. I ett senare steg blev utvidgningen till delstatsnivån permanentad, och kom att omfatta ämnesområdena läsning, skrivning, matematik och naturvetenskap för årskurserna 4 och 8, fortfarande dock på frivillig bas. I delstats NAEP använder man samma instrument som i nationella NAEP men urvalen är olika och administrationsprocedurerna skiljer sig delvis.

National Assessment Governing Board

Ett ytterligare förslag i Alexander och James (1987) rapporten var att ett nytt styrorgan för NAEP skulle inrättas. Detta skulle vara det policyformulerande organet, och bland annat fastställa ramverk, besluta om vilka ämnen och årskurser som skall utvärderas, och besluta om principer för rapportering av resultat. Anledningarna till att det fanns behov av ett brett förankrat policyskapande organ var att NAEP sedan 1984 genomfördes av den privata institutionen ETS, och att den administrativa styrningen blev förlagd till det statliga organet National Center for Education Statistics (NCES). År 1988 inrättades National Assessment Governing Board (NAGB) som en brett sammansatt styrelse, med representanter för båda partierna (i allmänhet på guvernörsnivå), skolledare, lärare, och lekmän (se Vinovskis, 1998). I NAGB ingår också en expert på pedagogiska mätningar (nu Andrew Porter och dessförinnan Edward Haertel).

År 1989 formulerade presidenten sex nationella utbildningsmål, och preciserade också inom vilka ämnesområden och årskurser som resultaten skulle utvärderas mot nationella och internationella kriterier. De angivna ämnesområdena och årskurserna matchade de som fokuserats inom NAEP, men genom formuleringen av utbildningsmål tillkom i NAEP en värdering av resultaten mot explicitgjorda ”performance standards”. Detta innebar att NAEP inte endast skulle beskriva elevers kunskaper och färdigheter, utan även värdera dessa i relation till vilka kunskaper och färdigheter elever bör besitta.

Det var NAGBs uppgift att utveckla principer för rapportering av NAEP-resultaten i termer av standards, och år 1990 fattade NAGB beslut om att som ett komplement till den vanliga rapporteringen även använda rapportering i form av andelen elever som nått olika på förhand definierade prestationsnivåer. Detta visade sig vara ett utomordentligt kontroversiellt beslut, som gav upphov till en stor mängd diskussioner och konflikter under hela 1990-talet, och som också visat sig få djupgående konsekvenser både för NAEP, och för utvecklingen av tänkande i termer av standards. Detta beskrivs närmare i kapitel 6.

Designdiskussioner under 1990-talet

Vid 1990-talets början hade i huvudsak NAEP antagit en komplex struktur, med de två huvudkomponenterna LTT NAEP och huvud NAEP, och där den senare komponenten i sin tur bestod dels av nationella NAEP, dels av delstats NAEP. Projektet byggde också på en mycket komplex teknik, som var kostsam och innebar långa fördröjningar av publiceringen av resultaten. Bland annat av dessa skäl tog NAGB vid 1990-talets mitt initiativ till en ”redesign” av NAEP. Utgångspunkten var en formulering av NAEPs primära mål, och ett antal delmål och rekommendationer. Bland annat önskade NAGB:

- Vidga omfattning, frekvens och systematik i genomförandet av utvärderingarna. Bland annat önskade man vidgning till flera ämnen: “History, geography, the arts, civics, foreign language, and economics also should be tested on a reliable basis according to a publicly released schedule adopted by NAGB”.

- Öka flexibiliteten i genomförande och rapportering, för att kunna få både snabbare resultatredovisning och fördjupad analys.
- Snabbare resultatpresentation, med en första rapportering inom 6 månader efter genomförd testning.
- Mer varierade rapporteringsformer, med såväl snabb rapportering efter varje utvärdering, som fördjupad analys av resultaten inom varje ämnesområde vart tionde år.
- Förenkling av den komplicerade statistiska teknik som hade introducerats år 1984.
- Transformera LTT NAEP och huvud NAEP till ett enda system för att mäta förändring over tid, i första hand genom att huvud NAEP används för trendmätning.
- Fortsatt rapportering av resultaten i termer av standards.
- Koppling av NAEP till internationella undersökningar.
- Förbättringar i datakvaliteten för åk 12.
- Aggregering av data från delstats NAEP för att få nationella NAEP resultat.
- Användning av modern teknologi för datainsamling och rapportering.
- Ramverk och testspecifikationer som är stabila under minst tio år.
- Beräkning av kostnadseffekter i samband med förändringar av ramverk och testspecifikationer.
- En blandning av flervalsuppgifter, uppgifter som kräver korta svar, och ”performance” uppgifter.
- Procedurer som gör det möjligt för delstater, skoldistrikt och andra att genomföra studier som utnyttjar NAEP.

I en närmast kontinuerligt pågående serie utredningar, undersökningar och utvärderingar har möjligheterna att införa dessa och andra förändringar utretts. Stora förändringar har också skett, men i många fall kvarstår den ursprungliga designen. Sammanläggningen av LTT NAEP och huvud NAEP har inte genomförts, och önskemålen om att de komplexa statistiska procedurerna skall förenklas har man inte kunnat tillmötesgå. Visserligen har flera ämnen än läsning, matematik och naturvetenskap undersökts, men det har i allmänhet skett inom specialstudier utan ambitioner att studera förändring över tid.

I många avseenden har sålunda NAEP visat sig vara påtagligt resistent mot förändringar. Detta återspeglar delvis det faktum att de kostnader och tekniska komplikationer som är förknippade med nationell utvärdering sätter gränser för vad som är genomförbart. Erfarenheterna från 1980-talets glada experimentlusta med olika procedurer har också lett till en återhållsamhet när det gäller att introducera förändringar som riskerar att påverka möjligheterna att göra jämförelser över tid.

Förändringar under 2000-talet

Vid 2000-talets början hade NAEP sålunda antagit en mycket komplex struktur, och bestod dels av nationella NAEP, dels av delstats NAEP. Nationella NAEP i sin tur består dels av en separat komponent för att bestämma långtidstrend (dvs LTT NAEP), dels huvud NAEP. Den senare omfattar nationellt representativa urval av elever i årskurserna

4, 8 och 12. Varje ämnesområde utvärderas två, tre eller fyra gånger under en 12-årsperiod, vilket gör det möjligt att studera förändring över tid. I nationella NAEP ingår också studier, som kallas *specialstudier*, och som inte är baserade på storskalig survey-teknik. Exempel på specialstudier är undersökningar av högläsning och mer omfattande skrivning. I specialstudierna har man inte ambitionen att studera förändringar i prestationer, och dessa studier omfattar i allmänhet ingående beskrivningar av läroplaner och undervisning inom de undersökta områdena.

Delstats NAEP omfattar representativa urval av elever från de delstater som valt att delta i utvärderingarna. Samma instrument används som i nationella NAEP, men med begränsning till ämnesområdena läsning, skrivning, matematik och naturvetenskap och till årskurserna 4 och 8.

I Bilag1 1 förtecknas alla de utvärderingar som genomförts under åren 1969 – 2004 inom de olika delarna av NAEP. Som framgår av Bilaga 1 är de olika ämnesområdena mycket ojämnt representerade i utvärderingarna. I nationella NAEP har sålunda läsning och matematik förekommit vartannat eller vart fjärde år.

Av Bilaga 1 framgår också att LTT NAEP studierna trots NAGBs ambitioner inte förenats med huvud NAEP, och som framgår av Bilaga 2 kommer LTT NAEP att fortsätta åtminstone till 2017. Under 2003 och 2004 genomgick emellertid LTT NAEP en modernisering för matematik och läsning. Tidigare hade man använt sig av den matris-samplingsmodell som användes i den ursprungliga NAEP-designen, och där det inte fanns något överlapp mellan olika block av uppgifter, men man lämnade nu denna till förmån för en BIB-design. Man lämnade också den bandspelarstyrda administrations-modellen och eleverna läser nu istället instruktioner och uppgiftsformuleringar.

Enligt Bushs “No Child Left Behind” lag som antogs år 2001 har NAEP ålagts en ny uppgift, nämligen att vara en kontrollinstans för de utvärderingssystem på elevnivå som delstaterna har ålagts att införa. Detta innebär en kraftig förändring av NAEPs roll i det att det får en tydlig plats inom det ansvarsutkrävande systemet. Konsekvenserna av detta är ännu oklara, men den innebär dels att den tidigare frivilligheten i deltagande i delstats NAEP upphör, dels att delstats NAEP skall genomföras vartannat år i matematik och läsning i åk 4 och 8. Vidare genomförs nu datainsamlingen inom delstats NAEP av anställd NAEP personal, och kostnaderna bärs av federala medel. En följd av detta är att budgeten för NAEP tredubblats, och den uppgår nu till över 100 miljoner USD/år.

Ytterligare en förändring under 2000-talet är att man nu i försöksform rapporterar resultat även för vissa större skoldistrikt.

Utvärderingar av NAEP

Under årtiondenas lopp har en stor mängd utvärderingar genomförts av NAEP. Vissa har beslutats av kongressen, andra har initierats av NAGB och NCES, Åter andra har genomförts i form av forskningsprojekt och ofta tillkommit på initiativ av rådgivande paneler av forskare vid National Academy of Education, National Reseach Council,

Educational Testing Service eller American Institutes for Research, för att endast nämna några.

I en genomgång av en del av dessa utvärderingar konstaterar Linn (2004):

> Although most of the evaluators have not hesitated to be critical of many aspects of NAEP, the general tenor of the evaluators' conclusions have been quite positive. NAEP has been subjected to close scrutiny and found to be a valuable source of information. (Linn, 2004, 291-292).

Den senaste och mest genomarbetade utvärderingen av NAEP har genomförts av en kommitté utsedd av National Research Council, och redovisats av Pellegrino, Jones och Mitchell (1999). En huvudslutsats är att NAEP:

> … has achieved prominence as the country's primary vehicle for monitoring levels of educational achievement. In fact, many groups want more NAEP – more often, more subjects, and with faster reporting – albeit at less cost. The popularity of the nation's national assessment program is a blessing, but also a curse: much of NAEP's current popularity is a product of these pressures, and its capacity for change may be limited by its prominence (s 10-11).

Utvärderingen argumenterar för att NAEP måste förenklas, och lämnar en rad förslag till hur NAEP kan förbättras. Jag återkommer längre fram till en diskussion av dessa förslag.

Lyle V. Jones har varit med under hela NAEPs utvecklingsperiod, och han (t ex Jones, 1999) är kritisk mot mycket av utvecklingen av NAEP, och menar att Tyler's ursprungliga idéer om en ”national assessment” har förfuskats:

> Over the past 30 years, some of Tyler's objectives for NAEP have survived, but just barely. First, the desired rich variety of exercises was compromised, in favor of more traditional multiple-choice and short-answer items, the kinds of items with which testing companies were familiar. Exercises became quite homogeneous in difficulty, with fewer very easy or very difficult ones. The young-adult sample was eliminated, and school grade has replaced grade as the primary unit of assessment. The ten subject areas have received uneven attention, with math, reading, science, and writing assessed far more often than literature, social studies, art, music, citizenship and career development. No longer are exercises read aloud, nor has an "I don't know" alternative been retained. For state assessments, local school personnel now administer the exercises, which raises questions about the uniformity of administration.
>
> Instead of reporting a percent-correct score for each exercise, scale scores were developed, for large clusters of exercises. More recently, reporting has been by "achievement levels," so as to compare actual performance with how good performance "should be".

> Using IRT technology, scores are now imputed for each child in the sample, even though different children take different sets of exercises. Imputed scores are then averaged for any specified subgroup of children.
>
> Many of these changes were well-intentioned, and some clearly are supported by psychometric considerations and by the need to better communicate results to the public. Nonetheless, some changes have compromised Tyler's vision of assessment. (Jones, 1999, s 4).

Jones (1996) noterar också att trycket på NAEP att bli ett ansvarsutkrävande system ökat över tid, och inte minst sedan man år 1994 tog bort förbudet mot att rapportera resultat på lägre nivå än delstatsnivå. Enligt Jones riskerar emellertid en utveckling i denna riktning för NAEP att få negativa konsekvenser för hela tanken att programmet skall fungera som en nationell indikator av nivån på kunskaper och färdigheter, inte minst därför att elevernas motivation i hög grad påverkar resultaten.

Jones (1996) betonar starkt vikten av att inte genomföra ändringar i systemet, därför att de förstör möjligheterna att studera förändringar över tid:

> The primary worth of NAEP has been as a monitor of changes in achievement for the nation and, more recently, for participating states. Thus NAEP has served as an indicator of educational achievement, much as the consumer price index and the nation's unemployment rate are indicators of economic conditions. In each case, comparability over time is of paramount importance to the interpretation of results. Only with extreme diligence can procedural stability be protected (Jones, 1996, s 20).

Jag återkommer även till dessa synpunkter i den fortsatta diskussionen.

3. Erfarenheter från andra nationella utvärderingssystem

Vid sidan NAEP finns en lång rad nationella utvärderingssystem och i detta kapitel beskrivs några av dessa.

Internationellt jämförande studier

Internationellt komparativa studier av elevers kunskaper och färdigheter har inte i första hand som syfte att vara nationella utvärderingssystem, men för många länder tjänar sådana studier samma funktioner som NAEP gör i USA. De internationella studierna står också inför likartade metodologiska utmaningar som de nationella utvärderingssystemen, nämligen att beskriva nivå och förändring över tid i kunskaper och färdigheter på utbildningssystemnivå.

IEA-undersökningarna

Ungefär samtidigt som planeringen av NAEP inleddes genomförde IEA (The International Association for Evaluation of Educational Achievement) de första komparativa studierna av elevers kunskaper och färdigheter i olika skolsystem, t ex First International Mathematics Study år 1964. Under de dryga 40 år som gått sedan den första studien har IEA genomfört ett stort antal jämförande studier av elevers kunskaper och färdigheter inom olika ämnesområden. IEA-organisation har vuxit till att ha 66 länder som deltagare, och med en världsomspännande organisation, bestående bl a av ett internationellt högkvarter i Haag, ett ”Data Processing Center” i Hamburg, ett ”International Study Center” i Boston, och en grupp urvalsexperter vid Statistics Canada.

Tyngdpunkten i IEAs nuvarande verksamhet ligger på genomförande av regelbundet återkommande studier dels av läsfärdighet (PIRLS, ”Progress in International Reading Literacy Study), dels av matematik och naturvetenskap (TIMSS, ”Trends in Mathematics and Science Study”). Antalet deltagande länder är ständigt växande, och i de studier som nu planeras förväntas 50 – 60 länder delta. Utöver dessa återkommande undersökningar pågår och planeras studier inom flera olika områden.

Bland IEA-studierna är det två som är av särskilt stort intresse i detta sammanhang, nämligen PIRLS och TIMSS. En anledning till detta är att dessa studier är planerade att upprepas i femårscykler för att göra det möjligt att studera förändringar i nivån av kunskaper och färdigheter över tid. En annan anledning är att dessa studier i många avseenden bygger på den metodologi som utvecklades för NAEP i samband med 1984 års redesign. Att de internationella studierna anammat de metoder som utvecklats inom NAEP har naturligtvis sin grund i att de metodologiska utmaningarna är likartade: syftet är att mäta och beskriva kunskaper och färdigheter inom breda domäner utan vägledning av en bestämd läroplan. Att IEA kom att anamma dessa metoder vid 1990-talets början har också sin grund i att en av centralpersonerna inom utvecklingen av NAEP, Albert Beaton, vid denna tidpunkt flyttade från ETS till Boston College, och där etablerade det

International Study Centre, varifrån först genomförandet av TIMSS 1995 (akronymen stod då för Third International Mathematics and Science Study) leddes och dit sedan ansvaret för genomförandet av de regelbundet återkommande TIMSS och PIRLS studierna förlagts. Även ett antal andra personer har antingen flyttat från ETS till Boston College, eller delat sin tid mellan de två institutionerna (t ex Ina Mullis, Jay Campbell, och Eugene Johnson).

Både TIMSS och PIRLS bygger på matrissamplingsdesign, MML-skattningar, och ”plausible values” för sekundäranalys. Vad som är särskilt intressant i detta sammanhang är att hela designen för trendanalys bygger på IRT-teknik och man tillåter att en relativt stor del av uppgifterna publiceras efter varje undersökningsomgång. Till skillnad från LTT NAEP använder man sig sålunda inte av samma uppsättning uppgifter för att mäta förändring över tid.

Programme for International Student Assessment (PISA)

År 2000 genomförde OECD den första omgången av PISA (Programme for International Student Assessment), i vilken samtliga OECD-länder och ett antal associerade länder (sammanlagt 41 länder) deltog i en undersökning av 15-åringars läsförståelse och kunskaper inom matematik och naturvetenskap. I denna undersökning var läsförståelse i fokus, medan matematik och naturvetenskap undersöktes med ett mer begränsat antal uppgifter. År 2003 upprepades PISA-undersökningen inom samma tre ämnesområden, men nu med fokus på matematik. Även fortsättningsvis kommer PISA att upprepas vart tredje år med fokusering på ett huvudområde i varje undersökning.

Även i PISA är huvudsyftet att studera förändring över tid. I varje treårscykel erhålls information om utvecklingen inom alla tre ämnesområdena, och vart nionde år erhålls fördjupad information om förändringen inom ett av de tre områdena. Denna design reser den intressanta frågan med vilken grad av säkerhet förändringar mäts dels över treårscyklerna, dels över nioårscyklerna.

Även inom PISA används matrissampling, MML-teknik och plausibla värden. Det finns dock vissa skillnader i den statistiska grundmodellen jämfört med den som används inom NAEP, PIRLS och TIMSS. Adams, Wilson & Wu (1997) har utvecklat en alternativ skattningsmetod för MML, som formulerats i termer av en tvånivåmodell. Enligt denna modell är individnivån den första nivån, där en IRT-modell i form av en generaliserad Rasch-modell specificerar relationer mellan observerade prestationer och strukturella karakteristika hos uppgifterna, och populationen av individer den andra nivån. Modellen tillåter direktskattning både av parametrarna i IRT-modellen och olika populations-parametrar. Formellt är denna modell ekvivalent med den modell som används i NAEP, men i NAEP används en tvåstegsteknik. Simuleringsstudier som genomförts av Adams, Wilson & Wu (1997) visar att denna direktskattningsmetod och NAEP-tekniken ger mycket likartade resultat. En annan skillnad är att man i PISA inte använder sig av information om bakgrundsvariabler i MML-skattningarna, medan man däremot använder sådan information vid beräkningen av plausibla värden (se Adams, 2002, s 107).

Ytterligare en viktig skillnad mellan PISA och de övriga utvärderingsprogrammen är att varje elev som deltar i PISA genomför uppgifter från alla tre ämnesområdena. Den statistiska modellen är också en flerdimensionell modell som tillåter skattning av relationer mellan prestationer inom de tre områdena.

Studier av vuxnas läskompetens

År 1985 genomfördes i USA the Young Adult Literacy Study (Kirsch & Jungeblut, 1986), med syfte att kartlägga läskompetensen bland vuxna. Denna undersökning har senare följts upp av flera undersökningar. En liknande serie internationellt komparativa undersökningar av vuxnas läsförståelse har genomförts av OECD (International Adult Literacy Study, IALS). Sammantaget 23 länder har deltagit i tre olika undersökningsomgångar mellan 1994 och 1998.

I dessa studier definieras läskompetens (”literacy”) brett och specificeras att omfatta tre områden: sammanhängande text, ”documents” (icke-kontinuerlig text, som tabeller, scheman, och kartor), och kvantitativ literacitet. Även här används matrissamplingsteknik och samma skattningsmetoder som inom NAEP.

Diskussion

Denna kortfattade översikt visar att den senaste generationen internationella studier anammat i huvudsak den metodologi som utvecklades för NAEP vid 1980-talets början vid ETS. Som redan påpekats har detta sin grund i att även de internationella studierna ställs inför uppgiften att avgränsa och strukturera breda domäner av kunskaper och färdigheter, utan att kunna förlita sig på en läroplan. Lösningen är då att genom en konsensusskapande process skapa ett ramverk, och sedan konstruera ett stort antal uppgifter som täcker av de domäner som identifieras i ramverket. Detta i sin tur kräver en matrissamplingsdesign för genomförandet av undersökningen, med ty åtföljande komplikationer i de statistiska operationer som krävs för valid inferens.

Det framstår dock som tydligt att denna teknologi, trots sin komplexitet, nu efter ca 20 år nått en sådan stabilitet och mognad att den används mer eller mindre rutinmässigt i komparativa sammanhang. Även för att studera förändring över tid verkar nu de metodologiska problemen som drabbade NAEP under lång tid åtminstone delvis ha bemästrats. Inom detta fält framstår det dock som om det fortfarande finns skillnader mellan olika tekniska lösningar, vilka finns anledning att ta upp till fördjupad diskussion.

Nederländerna

År 1986 påbörjades det holländska nationella utvärderingsprojektet PPON (”Periodieke Peiling van het Onderwijs Niveau”). Detta är ett ambitiöst europeiskt nationellt utvärderingssystem, och därför av särskilt intresse i detta sammanhang. Dokumentationen kring PPON föreligger dock praktiskt taget enbart på holländska, varför information

insamlats genom en studieresa till det holländska provinstitutet CITO som har ansvaret för att genomföra PPON (Söderberg, 2005).

Det uttalade motivet bakom införandet av PPON var att ge alla aktörer i det starkt decentraliserade holländska skolsystemet (inklusive inspektoratet) en empirisk grund för sitt agerande. Detta innebar att CITO skulle svara för instrumentkonstruktion, datainsamling och resultatrapportering, men däremot inte för värdering av resultat, policybetonade slutsatser eller rekommendationer. Det senare förutsattes ske på nationell politisk nivå och i professionella kretsar.

PPONs undersökningar har gjorts bland tolvåringar (åk 8) och i vissa fall bland nioåringar (åk 5) inom följande ämnen:

- Aritmetik som i sin tur består av 22 olika komponenter i åk 8 och 13 olika komponenter i åk 5. Andra aspekter som t.ex. algebra eller geometri ingår inte i primärutbildningen. Mätningar har gjorts fyra gånger – 1987, 1992, 1997 och 2004.
- Modersmål har studerats i fyra aspekter: Läsförståelse, hörförståelse, skrivning och tal. Med start 1989 gjordes dessa som mätningar av delkomponenter vid fyra tillfällen fram till 2000, då planen reviderades så att varje delkompetens skulle mätas vart fjärde år.
- Engelska har undersökts två gånger, 1991 och 1996, och omfattade bl. a. läsförståelse, ordförråd och uttal. En tredje mätning planeras till 2006.
- Undersökningar har också gjorts av ”Social and natural sciences”. ”Social sciences” består i Nederländerna av historia (inklusive nutidshistoria, vilket är ungefär detsamma som ”civics”, dvs. samhällskunskap) och geografi. ”Science” består av biologi, fysik och teknik. Fram till 1995 mättes historia, geografi, biologi och fysik samlat två gånger (1991 och 1995) men från år 2000 har de av ekonomiska skäl mätts separat olika år (historia 2000, geografi 2001, biologi 2002…).
- Musik har undersökts en gång, 1997, och omfattade bl.a. sång (individuellt och i grupp); formella tester om instrument, länder, kulturer, olika typer av orkestrar och notering; och praktiska tester, som att repetera en rytm på ett träblock. Musikutvärderingen var delvis influerad av uppgifter som utvecklades för det svenska NU-92 projektet.
- Idrott har undersökts en gång, 1994, och omfattade prestationer i ca 25 traditionella gymnastik- och idrottsmoment som hoppa bock, springa 60 m, höjdhopp, kullerbytta, kast med liten boll osv.
- Bild har ingått en gång 1996, och omfattade såväl konstförståelse som tre praktiska ”teckningstester” – avbildning av ett föremål, illustrera en uppläst berättelse; samt designa en 100 Euro-sedel.
- Härutöver har vid ett tillfälle en mindre undersökning av trafikkunskap genomförts. Denna genomfördes som ett praktiskt ”körprov” (t.ex. att följa trafikregler då man cyklar) på en lekplats dit barn från olika skolor fraktades.

En matrisbaserad design utnyttjas för att bjuda uppgifter i de återkommande delarna av PPON, dvs holländska samt matematik. Resultaten i dessa delar analyseras på skalnivå. Dessa skalor utnyttjar IRT-baserad programvara som utvecklats vid CITO (OPLM, One Parameter Logistic Model).

I sammanfattning kan PPON sägas ha genomgått fyra utvecklingsfaser. Den första fasen inleddes med undersökningen av aritmetik bland 12-åringar och utvidgades sedan att omfatta fler ämnen och upprepade mätningar. En andra fas inleddes 1997, då mätningarna fokuserades på aritmetik och modersmål. Bakgrunden var de dåliga holländska resultaten i IEA Reading Literacy 1991. År 2003 inträffade en tredje, kortlivad, fas då mätningar i flera ämnen återupptogs eller planerades. Den senaste tiden har dock systemets legitimitet återigen alltmer minskat, särskilt efter att autonoma skolor har blivit allt vanligare och offentlig redovisning av aggregerade resultat på andra prov som genomförs på individnivå blivit vanliga.

Det tycks finns flera olika anledningar till den sviktande legitimiteten. Många avnämare är mer intresserade av den mer ”high-stake” betonade information som erhålls genom aggregering av individuella resultat, och som då bland annat kan redovisas på skolnivå. Här erhålls dock ingen information om trender i den nationella utvecklingen.

Inte heller PPON synes emellertid ha varit särskilt framgångsrikt i att beskriva de trendmässiga förändringarna i kunskaper och färdigheter. En anledning till detta är den brist på systematik och konsekvens i upprepningen av mätningarna över tid som framgår av ovanstående beskrivning. Redovisningen av resultaten av trendanalyserna framstår inte heller som lämpligt utformad då syftet är att ge en övergripande beskrivning. För aritmetik redovisas exempelvis andelen elever som uppnår olika ”nivåer” för var och en av det stora antalet olika komponenterna, varvid trenden framgår. De grafiska kurvor som visas för de olika komponenterna ger dock ett mycket motsägelsefullt intryck och det är svårt att få klart för sig i vilken riktning som holländska barns kunskaper i aritmetik har utvecklats. Huvudanledning till detta är den långt drivna nedbrytningen av ämnet i olika komponenter, vilket grundas på att man upprätthåller strikta krav på endimensionalitet i mätningarna.

En ytterligare förklaring till det låga intresset är att det finns stort behov av förklaringar till funna resultat och diskussion kring möjliga handlingsalternativ. PPON har dock inte som huvudsyfte att ge underlag för mer fördjupade analyser av orsakerna till de erhållna resultaten. Detta innebär att resultaten är svåra att både tolka och använda för såväl politiker som lärare. Detta leder i sin tur till metodproblem i form av bortfall, som ytterligare undergräver systemets funktionalitet och legitimitet (se van Lent och Bakker, 2004).

England

Även England har vad som kan betraktas som ett nationellt utvärderingssystem, men det är inte urvalsbaserad utan individfokuserat. Basen för det engelska systemet utgörs av en omfattande nationell läroplan med preciserade målbeskrivningar. Utvärderingssystemet

för 7-åringar introducerades år 1991, för 11-åringar introducerades prov i engelska, matematik och naturvetenskap år 1993, och för 14-åringar började man med prov i samma ämne år 1994. Alla elever som går i offentligt finansierade skolor i England, Wales och Nordirland genomgår dessa prov. Ett syfte med proven är att de skall vara diagnostiska och visa på olika elevers starka och svaga sidor. Resultaten sammanställs dock också i syfte att visa resultat på skol- och distriktsnivå.

Läroplanen är strukturerad i 8 olika prestationsnivåer som omfattar åldrarna 7 till 14 år. Samma definitioner av prestationsnivåerna används för alla åldersgrupper, men den procentuella fördelningen av eleverna över de olika nivåerna förväntas vara olika för de olika åldersgrupperna. Varje skola förväntas sätta upp mål för hur stor andel av eleverna som skall nå de olika nivåerna, och utvärderingen sker mot dessa mål. Det finns också nationella mål för hur stor andel av eleverna som skall nå de olika prestationsnivåerna.

Det engelska systemet har mött mycket kritik och har ifrågasatts av flera olika grupper. I synnerhet gäller detta proven för 7-åringar (van Lent & Bakker, 2004).

Nya Zeeland

År 1995 startade National Education Monitoring Project (NEMP; se Flockton & Crocks, 1999; Flockton, 1999) efter ungefär 30 års utredningar som hade identifierat behov av ”… regular, dependable and consistent information about the educational achievements, attitudes and interests of New Zealand students.” (Flockton & Crocks, 1999, s 1). Före 1995 fanns i Nya Zealand inget nationellt utvärderingssystem, utan man hade helt förlitat sig på deltagande i IEA-undersökningar. Det fanns dock behov att komplettera dessa eftersom de endast avsåg begränsade delar av läroplanen, och med begränsad variation i uppgifts- och svarstyper.

Som påpekats av Jones har många av Tyler’s grundläggande idéer övergivits inom NAEP, medan däremot ”… Tyler’s vision has materialized fully in New Zealand’s promising National Education Monitoring Project.” (Jones, 1999, s 4).

Ett uttalat syfte är att NEMP skall beskriva förändringar i kunskaper och färdigheter över tid, och att det skall ge information som är användbar både för formulering av policy och för praktisk verksamhet: ”An important goal of the project is to help identify what is being done well, areas of concern, and priorities for future improvement in student achievement” (Flockton & Crooks, 1999, s 2).

NEMP genomförs i en fyraårscykel, och två sådana cykler har fullbordats. Under varje fyraårsperiod utvärderas 15 olika ämnesområden:

- 1995/1999/2003: naturvetenskap; bild (”art”); och informationsfärdigheter (grafer, tabeller och kartor.
- 1996/2000/2004: läsa och tala; teknik; och musik.
- 1997/2001/2005: matematik; samhällskunskap; och informationsfärdigheter (bibliotek och informationssökning).

- 1998/2002/2006: skriva, läsa och lyssna; hälsa och idrott.

Utvärderingen genomförs dels i åk 4 (8-9 år; skolstart på Nya Zeeland är normalt vid 5 års ålder), dels i åk 8 (12-13 år). Varje år görs ett urval av 1 440 elever i åk 4 från 130 slumpvis valda skolor, och lika många elever i åk 8 från lika många skolor. Från varje skola väljs 12 elever slumpmässigt ut, och delas sedan upp i tre grupper med fyra elever i varje grupp. Varje sådan grupp av elever genomför sedan uppgifter från alla de ämnesområden som utvärderas det aktuella året. De tre grupperna av elever gör olika uppgifter, så NEMP använder sig av en matrissamplingsdesign som liknade den som användes i NAEPs ursprungliga utformning. Under en vecka arbetar varje elev i ungefär fyra timmar med NEMP-uppgifterna. Eleverna och deras föräldrar väljer själva om de skall delta i utvärderingen eller inte.

En grundtanke i NEMP är att eleverna skall kunna visa sina kunskaper och färdigheter inom ett ämnesområde, utan att resultaten påverkas av svagheter eller styrkor inom andra områden. Då man använder papper-och-penna prov riskerar läs- och skrivförmågan att påverka resultaten inom exempelvis naturvetenskap. Huvuddelen av uppgifterna presenteras därför muntligt av lärare, med video, eller på dator. Eleverna svarar vanligtvis muntligt, eller genom något praktiskt utförande, snarare än genom att skriva. Då uppgifterna avser något annat än att läsa och skriva hjälper läraren gärna eleverna att förstå texter och att kommunicera svaren. Eftersom varje grupp om fyra elever har tillgång till en lärare är också elevernas möjligheter att få den hjälp de behöver goda.

Den lärare som leder genomförandet är inte elevernas reguljära lärare, utan en annan lärare som i allmänhet kommer från den region där skolan ligger. De lärare som leder genomförandet av NEMP-uppgifterna har valts ut bland en större grupp sökande, och har fått en veckas utbildning för detta. De genomför sedan utvärderingar med 60 elever under en period om 5 veckor, och arbetar då två och två. NEMP-projektet betalar den skola där lärarna normalt har sin anställning för en vikarie. Man eftersträvar att varje lärare endast deltar som NEMP-lärare ett år, eftersom denna verksamhet betraktas som kompetens-utveckling. Enligt de utvärderingar som genomförts är detta också en mycket uppskattad och framgångsrik form av kompetensutveckling (Gilmore, 1999).

Man använder sig av fyra olika sätt att administrera uppgifterna:

- *Intervju med en lärare och en elev*: Varje elev arbetar individuellt tillsammans med en lärare, och hela sessionen spelas in på video.
- *Stationer*: Fyra elever, som arbetar oberoende av varandra, rör sig mellan fyra stationer, där de genomför olika uppgifter. Denna session spelas inte in.
- *Lag*: Fyra elever arbetar tillsammans för att lösa uppgifter. Denna session spelas in på video.
- *Oberoende arbete*: De fyra eleverna arbetar individuellt med papper-och-penna uppgifter.

En av anledningarna att stora delar av elevernas arbete spelas in på video är att all rättning och bedömning av elevernas svar sker efter att allt material samlats in. En del av

elevernas svar bedöms sålunda från videoinspelningarna, och annat från elevernas skriftliga svar eller annan produktion.

De bedömningar som kan göras enligt enkla regler genomförs av lärarstuderande, som i allmänhet har genomgått två till fyra år av sin lärarutbildning. Mer krävande bedömningar genomförs av verksamma lärare, som rekryteras från hela landet. Bedömningsprocessen inleds med genomgång av bedömningsanvisningarna, och ett antal exempel. Under bedömningsarbetet kontrolleras också graden av konsistens mellan olika bedömare.

För att utveckla ramverk och medverka vid konstruktion och urval av uppgifter använde man sig av 9 paneler, som sammantaget bestod av 65 läroplansexperter och praktiskt verksamma lärare. Huvuduppgiften för dessa var att identifiera de primära utfall som utvärderingen skulle omfatta, och då inte endast kunskaper och färdigheter utan även attityder.

Uppgifterna har konstruerats så att de uppvisar en hög grad av variation i svårighetsgrad, med både mycket lätta och mycket svåra uppgifter. Vissa uppgifter består av olika delar, med en inbyggd progression. Ambitionen har också varit att skapa uppgifter som intresserar eleverna och som får dem att göra sitt bästa. Den omfattande användningen av olika typer av praktiska ”hands-on” uppgifter har uppmärksammats som en positiv faktor för elevernas motivation. Presentation av uppgifter på dator och video gör det också möjligt att använda rikare och mer stimulerande material, samtidigt som sådana tekniker standardiserar presentationen.

Uppgifterna är i stor utsträckning gemensamma för åk 4 och åk 8, men vissa har ändrats så att de är mer åldersanpassade, och andra uppgifter förekommer endast för den ena åldersgruppen.

Efter varje session får eleverna värdera i vilken grad de tyckte om att arbeta med uppgifterna. Resultaten visar att i synnerhet eleverna i åk 4 har varit mycket positivt inställda till uppgifterna, men även eleverna i åk 8 har i stor utsträckning givit uttryck för en positiv inställning. Även elevernas föräldrar och deras lärare har reagerat positivt på uppgifterna och formerna för genomförandet av utvärderingen.

Analys och redovisning av resultaten sker på uppgiftsnivå, därför att ”… statistically aggregated data … can be largely meaningless and easily misused.” (Flockton & Crooks, 1999, s 4). För att denna redovisningsmodell skall vara meningsfull krävs att uppgifterna är kända, och efter varje tillfälle publiceras ungefär 2/3 av uppgifterna, medan återstoden återanvänds i nästa utvärderingscykel för att studera förändring över tid. Rapportering görs också för olika undergrupper, definierade av 9 variabler (t ex kön, elevens etnicitet, skolans socioekonomiska status, skolstorlek, region, och skolans etniska sammansättning). Rapporterna skrivs på ett sådant sätt att de skall vara tillgängliga för en bred publik, och de sprids i stor upplaga, bland annat till samtliga skolor på Nya Zeeland. En serie fördjupade analyser och forskningsrapporter publiceras också.

Ett utmärkande drag hos NAEP är man ser det som viktigt att involvera yrkesverksamma lärare i projektet. Ett skäl för detta är att man vill utnyttja den kompetens lärarna besitter:

> … teachers through their professional training, development and sustained experiences in the real world of the classroom have valuable and valid insights about the nature and appropriateness of curriculum in respect to the diversity of students they teach and the communities they serve… The teacher's view of schooling and curriculum is more than mere perception – it is highly adviced through practical reality. National monitoring accepts that any programme of assessment should not only acknowledge such experience, but also draw upon it. (Flockton, 1999, s 20).

Det andra skälet är att lärarna måste vara delaktiga för att NEMP skall kunna förbättra undervisningen:

> … a strong sense of commitment and ownership is fundamental to achieving genuinely meaningful professional development, and improvement of teaching and learning. Influence towards good practice is much more likely to arise from what professionals perceive to be credible, sensible and worthwhile. If they themselves are partners in the guided development of policy and practice, they are more likely to understand, support and accept its influence and direction. (Flockton, 1999, s 20).

Som redan nämnts innebär denna grundsyn att stora delar av arbetet inom NEMP genomförs av yrkesverksamma lärare. Framförallt gäller detta utveckling av uppgifter, administration av uppgifter, och bedömning av elevsvar.

Det är intressant att notera att NEMP i så stor utsträckning är troget mot de idéer som Tyler lanserade i utvecklingsarbetet med NAEP. Projektet synes också ha fått ett gott mottagande på Nya Zeeland. Det har också utvärderats av internationella forskare vid två tillfällen. Paul Black, Michael Kane och Robert Linn drog i en utvärdering som genomfördes 1996 slutsatsen att "… the project has considerable potential for advancing the understanding of and public debate about the educational achievement of New Zealand students. It may also serve as a model for national and/or state monitoring in other countries" (citerat från NEMP rapport 21). En ytterligare utvärdering genomfördes av Elliott Eisner, Caroline Gipps och Wynne Harlen år 1998, och även dessa utvärderare gav uttryck för uppskattning av NEMP, och gav också " … very helpful comments for further refinements and investigations." (citerat från NEMP rapport 21).

Diskussion

I detta kapitel har några ytterligare exempel på nationella utvärderingssystem beskrivits, och det skulle vara möjligt att ta upp även andra exempel (t ex Canada, Portugal, Skottland). Dock torde den variation som finns i huvudsak representeras av dessa exempel.

Det är slående hur olika de nationella utvärderingssystemen är utformade, men i huvudsak kan de ordnas längs en dimension där den ena extremen utgörs av anvars-utkrävande system och den andra extremen utgörs av utvecklingsorienterade system. De ansvarsutkrävande systemen, där det engelska utgör ett exempel, genomförs av alla elever, och har direkta konsekvenser för elever, lärare, skolor och skoldistrikt. En begränsad del av skolans ämnen ingår i utvärderingen och inom dessa fokuseras det mätbara. De utvecklingsorienterade systemen, där NEMP utgör det främsta exemplet, är urvalsbaserade och redovisar huvudsakligen resultat på nationell nivå, och för olika demografiska undergrupper. Ambitionen är att täcka av hela bredden av ämnen och att använda sådana uppgifter och former för utvärderingen att hela spektrum av undervisningsmål täcks in.

Denna variation återspeglar givetvis det faktum att de olika systemen har delvis olika explicita syften, men också att de genomförs i olika sociala och politiska kontext. Det finns dock anledning att särskilt ta upp till diskussion de olika syften som förknippats med nationella utvärderingsprojekt.

Det primära syftet för flera av projekten är att de skall ge en beskrivning på systemnivå av nivån på kunskaper och färdigheter, och hur denna förändras över tid. Enligt the NAEP Guide har detta program två huvudsyften: ”to reflect current educational and assessment practices and to measure change reliably over time.” (U. S. Department of Education, 1999, s 3). NAEP skall sålunda i första hand *beskriva* utbildningsresultat, och hur dessa förändras över tid.

Även om det är enkelt att konstatera att NAEP och de övriga nationella utvärderings-systemen har som ett huvudsyfte att beskriva utbildningsresultat på systemnivå kan detta göras på många olika sätt. Vad som framstår som en avgörande skillnad mellan olika utvärderingssystem är i vilken utsträckning man strävar efter *validitet* i beskrivningarna. I detta sammanhang är det två validitetsaspekter som är centrala. Den ena är innehålls-validitet, och som avser i vilken utsträckning utvärderingssystemet täcker av alla skolans ämnesområden och hela spektrum av mål för skolans verksamhet. NEMP har, exempel-vis, en hög ambitionsnivå vad gäller utvärderingens innehållsvaliditet, och även för NAEP är den uttalade ambitionen vad gäller innehållsvaliditet hög, även om det här skett en förskjutning över tid. I det engelska systemet är däremot innehållsvaliditet inte en högt prioriterad egenskap, utan fokus ligger här på en begränsad del av skolans ämnen och på begränsade delar inom dessa. Den andra validitetsaspekten är det som Messick (1989) benämner konsekventiell validitet, och som avser hur mätsystem påverkar våra värden och sätt att tänka, och hur de påverkar utformning och genomförande av olika verksamheter, som undervisning. En lång rad undersökningar har visat att utvärderings-system kan ha starkt inflytande på hur exempelvis lärare och elever definierar vad som är viktigt och vad som är mindre viktigt, och det är uppenbart att man i många utvärderings-system har som ett mer eller mindre uttalat syfte att genom utformningen av uppgifter påverka skolans sätt att arbeta i en positiv riktning. Detta gäller NAEP, och det gäller i än högre grad NEMP. Den konsekventiella validiteten är beroende av innehållsvaliditeten, men den är också i hög grad påverkad av i vilken utsträckning utvärderingen är ”low-stake” eller ”high-stake”. I det engelska systemet, som är ett ”high-stake” system kan

sålunda den låga graden av innehållsvaliditet förväntas medföra en stark negativ konsekventiell validitet, vilket också framgår av de intensiva diskussionerna kring detta system. Ett ”low-stake” system som NEMP riskerar, trots god innehållsvaliditet, att inte ha någon påverkan på utbildningssystemet, om man inte vidtar speciella åtgärder för att åstadkomma detta.

Även om beskrivning av nivå och trend är huvudsyften som det finns stor enighet kring, framgår det också av de intensiva diskussioner som förts kring exempelvis NAEP att många menar att det finns ett ytterligare mer överordnat syfte, nämligen att utvärderingarna skall *förbättra* utbildningssystemet. Ett problem med de extremt abstraherade beskrivningar som utvärderingarna ofta resulterar i är att de inte ses som intressanta och relevanta ur verksamhetssynpunkt, och många gånger har argumentet framförts att NAEP borde läggas ner av detta skäl.

Argumentationen kring utvärderingsresultat som en bas för att förbättra verksamheten förs enligt åtminstone tre helt olika argumentationslinjer. Enligt den första argumentationslinjen måste resultaten presenteras på en adekvat aggregationsnivå för att vara användbara för att kunna påverka utbildningssystemet. I ett decentraliserat utbildningssystem innebär detta att ett nationellt utvärderingssystem riskerar att presentera resultat på en för hög aggregationsnivå och argumenten för att införa delstats-NAEP var just att det är på delstatsnivån som viktiga beslut om finansiering och utformning av utbildningssystemet fattas. I England ligger mycket av beslutsmakt på skolnivå och hos elever och föräldrar genom att man där har ett fritt skolval. Det är därför naturligt att man där redovisar utvärderingsresultaten ner på skolnivå. Enligt denna argumentationslinje handlar sålunda möjligheten till förbättring om att redovisningen av resultat skall ske på rätt aggregationsnivå, men man lämnar också över till motsvarande beslutsnivå att identifiera orsaker och vidta åtgärder. Denna tilltro värdet av kvantitativ information har djupa historiska rötter:

> Nineteenth-century reformers had an abiding faith that the compilation and display of numerical data not only would reveal the inherent regularities in behavior, but also would suggest possible options for making changes. They believed that if policymakers and the public were presented with the appropriate comparative data on social reforms such as education, they would soon want to improve their own policies accordingly. (Vinovskis, 1998, s 3).

Enligt denna argumentationslinje sker sålunda förbättringen av skolan genom att de beskrivande utvärderingsresultaten bidrar till ett förbättrat beslutsunderlag för föräldrar, elever, lärare, skolledare, och lokala politiker.

Den andra argumentationslinjen betonar utvärderingens roll som förebild för verksamheten i skolan, och bland de utvärderingssystem som presenteras har företräds denna tydligast av NEMP:

> To produce or use assessment information solely for the purposes of accountability without connecting it to purposes and processes of improvement is

> to deny the very essence of educational assessment. National monitoring recognizes the value and impact of 'learning-integrated' assessment: that is, assessment which has the potential to benefit both teaching and learning while also giving clear and dependable information about student performance and achievement. (Flockton, 1999, s 12).

Den tredje argumentationslinjen betonar utvärderingarnas funktion för att nå fram till bättre förklaringar av vilka faktorer som är betydelsefulla för goda pedagogiska resultat. De insamlade data ses här som en resurs för fördjupad analys och forskning, som kan leda fram till resultat kring betydelsen av olika policyrelevanta faktorer. Möjligheterna att använda utvärderingsresultat på detta sätt för att åstadkomma förbättringar har varit mycket omdiskuterat, inte minst i anslutning till NAEP (se, t ex Pellegrino, et al., 1999) och de internationella undersökningarna (se t ex Härnqvist, 1975).

Vid sidan av beskrivning och förbättring av utbildning kan ett tredje huvudsyfte med nationella utvärderingssystem identifieras, nämligen *värdering* av i vilken utsträckning de resultat som uppnås skall betraktas som goda eller mindre goda. Detta syfte tillkom för NAEPs del vid början av 1990-talet, då standardsbaserad rapportering infördes. Värdering av resultat utgör också ett centralt inslag i det engelska systemet, och i PPON, men lyser med sin frånvaro i NEMP.

Denna diskussion leder fram till slutsatsen att åtminstone en del av skillnaderna i utformningen av de nationella utvärderingssystemen har sin grund i att man fäster olika avseenden vid de tre huvudsyftena beskrivning, förbättring, och värdering, liksom att innebörden i dessa varierar högst betydligt över de olika systemen. I nästa steg diskuteras därför vart och ett av dessa tre syften närmare.

4. Beskrivning av kunskaper och färdigheter

Som konstaterats ovan är ett dominerande huvudsyfte med NAEP och de övriga nationella utvärderingssystemen att beskriva nivån av kunskaper och färdigeter på nationell nivå, och då särskilt att studera förändringar över tid. De utvärderingar som gjorts av NAEP visar också att projektet varit framgångsrikt i att nå detta syfte. Av detta kan man dock inte dra slutsatsen att det är en enkel uppgift att beskriva kunskaper och färdigheter på nationell nivå, och man kan inte heller dra slutsatsen att detta är ett okontroversiellt företag.

Problemet att fastställa nivåer av kunskaper och färdigheter är ett mät- och generaliseringsproblem, vilket gör det naturligt att strukturera diskussionen kring de centrala begreppen i mät- och designläran:

- *Validitet*, varvid det vida validitetsbegreppet enligt Messick (1989) är en naturlig utgångspunkt. Enligt Messick formuleringar avser validitet inte endast giltigheten i de tolkningar som görs på grundval av mätningarna utan även de konsekvenser mätningarna har för individer och grupper, och för vårt sätt att tänka kring de undersökta fenomenen.
- *Reliabilitet*, eller tillförlitlighet.
- *Generaliserbarhet*.
- *Mätning av förändring*. Visserligen kan problemet att mäta förändring analyseras i termer av validitet, reliabilitet och generaliserbarhet, men eftersom detta problem bjuder på en lång rad speciella utmaningar, och eftersom det dessutom är ett speciellt syfte inom NAEP att mäta förändring tas detta upp till diskussion under egen rubrik.

I allmänhet är det inte svårt att åtminstone i princip nå långt när det gäller att uppfylla de krav som dessa metodbegrepp implicerar, men praktiska omständigheter och kostnader sätter bestämda gränser för vad som kan göras, vilket innebär att det praktiskt taget alltid handlar om att göra optimeringar och avvägningar mellan olika lösningar. Sådana avvägningar diskuteras nedan.

Validitet

Champagne och Pearson (2003) beskriver de reaktioner som vanligen följer på publicering av NAEPs resultat och exempeluppgifter:

> Teachers and school-based subject matter coordinators criticize mathematics and science tests, claiming that their students understand the mathematics and science contained in the released items, but that the reading demands are so great that they cannot perform well despite their understanding of the content domain. Teachers and coordinators level similar criticism against items requiring extended responses. Teachers and coordinators claim that the students understand these items but cannot express that understanding in written form. Ironically, then, two

> other domains of NAEP assessment, reading and writing, may be interfering with our capacity to assess mathematics and science with high degrees of validity.
>
> Representatives of teacher and coordinator professional societies, argue that many of our tests, including not only mathematics and science but also reading and writing assessments, do not represent the subject matter content valued by educators. … Often the criticisms focus on what these critics claim is the over-representation of items measuring lower level information and the under-representation of items measuring higher level cognitive abilities, such as problem solving or inquiry.
>
> These same educators are critical of the alignment of content on NAEP with student's opportunity to learn, claiming that the subject domain sampled by NAEP assessments does not correspond with the requirements of state standards …
>
> Representatives of the academic disciplines criticize the choice of principles tested, claiming that they do not represent the most powerful or newest ideas of the discipline. Discipline-based critics claim that multiple-choice items do not assess true understanding. They are also highly critical of the accuracy of the items, pointing out, for example, that a response scored as correct may not be *exactly* correct in the context described in the stem of the item. (Champagne & Pearson, 2003, s 5).

Denna kritik reser en hel rad frågor om på vilket sätt nationella utvärderingssystem bör utformas för att på ett rättvisande sätt beskriva elevernas kunskaper och färdigheter, i synnerhet som synpunkterna från de olika intressenterna ofta leder till oförenliga krav på hur uppgifterna skall utformas.

Innan denna fråga tas upp till mer utförlig diskussion finns det dock anledning att peka på ett annat problem, nämligen vilka ämnesområden som skall omfattas av det nationella utvärderingssystemet. Enligt de ursprungliga planerna skulle NAEP omfatta i stort sett samtliga skolans ämnesområden, vilka skulle undersökas enligt ett rullande schema. Förteckningen över de utvärderingar som genomförts inom NAEP i Bilaga 1 visar dock på en betydande slagsida till förmån för läsning och matematik, liksom för naturvetenskap. Visserligen har utvärderingar genomförts inom de flesta ämnesområden någon gång sedan 1969, men det har ofta skett inom NAEPs program för specialstudier, där det inte finns någon ambition att studera förändring över tid. En analys av vilka områden som omfattas av NEMP pekar på en betydligt bättre täckning av vad som behandlas i skolan, men här kan man istället notera att områdena läsning, matematik och naturvetenskap har en relativt begränsad förekomst. Av totalt 356 uppgifter i åk 8 under den första fyraårscykeln var sålunda endast 17 läsuppgifter. Inom matematik förekom 46 uppgifter i åk 8, men de avsåg endast området ”numeracy skills”. En möjlig anledning till detta kan vara att Nya Zeeland inom dessa områden förlitar sig på de internationella studierna som en ytterligare informationskälla.

Ett av skälen till varför områden som läsning och matematik är överrepresenterade i nationella utvärderingssystem är givetvis att dessa är betydelsefulla områden, som ges stort utrymme i skolans verksamhet. Men det finns säkert andra skäl, som att det inom

dessa områden finns en lång tradition av kunskaps- och färdighetsmätning, medan det inom andra områden inte finns en sådan tradition. Befintlig kompetens har sålunda säkerligen stor betydelse för vilka områden som prioriteras. Det faktum att utvärderingarna i stor utsträckning genomförs med papper-och-penna instrument i klassrumskontext sätter också restriktioner på vilken typ av kompetenser som kan fångas upp och då i synnerhet inom olika färdighetsområden. Inom vissa områden finns också motstånd mot utvärdering:

> Within the Arts, for example, there is a sectional lobby which is suspicious and resentful of the very idea of assessment. To add to the challenge, very few examples are available of suitably valid assessment tasks which offer good models as starting points. (Flockton, 1999, s 6).

Begränsningarna i möjligheterna att genomföra utvärderingar inom alla skolans fält med de mer traditionella metoderna för att mäta kunskaper och färdigheter kan sålunda medföra att centrala delar av skolans arbete inte uppmärksammas i de nationella utvärderingarna, vilket måste betecknas som ett (innehålls)validitetsproblem. Detta problem behöver uppmärksammas i samband med planeringen av ett nytt nationellt utvärderingssystem.

Givet ett visst ämnesområde kan validitetsproblemet delas upp ett antal delproblem:

- Ramverkets validitet;
- uppgiftspoolens validitet i förhållande till ramverket; samt
- validiteten i de använda uppgiftstyperna och metoderna för att bedöma elevernas svar.

Ramverkets validitet

Det fundament som utvärderingen av ett ämnesområde vilar på kallas ”ramverk” (”framework” på engelska). Ett ramverk definierar huvudstrukturen av kunskaper och färdigheter inom ett ämnesområde, anger vilka typer av uppgifter som skall användas, hur stor andel av det totala antalet uppgifter en viss uppgiftstyp skall utgöra, osv. I det amerikanska skolsystemet finns ingen gemensam läroplan att utgå från då ramverket skall konstrueras. Detta innebär att de ramverk som NAEP bygger på skapas genom förhandlingsprocesser där olika intressenter i utvärderingarna är involverade, tillsammans med ämnesexperter, lärare och mätexperter. I allmänhet sätts en panel samman för varje ramverk som skall byggas. Panelmedlemmarna väljs på så sätt att de skall representera skilda politiska och pedagogiska synsätt, liksom olika grupper i befolkningen.

Avsaknad av en gemensam läroplan innebär att ramverken måste vara breda nog att kunna inrymma hela mångfalden av innehållsval och sätt att lägga upp undervisningen. Men även i de fall då det finns en läroplan finns det skäl att inte binda ramverket så hårt till denna att ramverket måste ändras så snart läroplanen ändras. Inom NEMP spelade läroplanen stor roll vid utvecklingen av ramverken ”… but without attempting to slavishly follow the finer details of current curriculum statements. Such details change

from time to time, whereas national monitoring needs to take a long term perspective if it is to achieve its goals." (NEMP report 21, s 6).

Sedan 1988 har NAGB ansvaret för att utveckla NAEPs ramverk, och man har publicerat ramverk för alla de utvärderingar som genomförts sedan 1990. I allmänhet görs en smärre förändringar av ramverket inom ett område från en utvärdering till en annan, och avsikten är att ett ramverk i huvudsak skall vara användbart under minst 10 år. Anledningen till detta är att förändringar av ett ramverk hotar att omöjliggöra studier av förändring över tid.

Som redan nämnts hade den första generationen av NAEP-instrument ambitionen att minska andelen flervalsuppgifter till förmån för uppgiftstyper som kräver producerade elevsvar. Ambitionen att ha en relativt stor andel öppna uppgifter har hela tiden funnits i NAEP, och då kraven under 1980- och 1990-talen ökade på uppgifters grad av autenticitet har också ambitionsnivån vad gäller uppgifters komplexitetsgrad och öppenhet höjts inom NAEP. Genom att urval görs av både elever och uppgifter är det inom NAEP genomförbart att använda relativt stora andelar öppna uppgifter, även om det naturligtvis innebär högre kostnader än flervalsuppgifter, och även har stor betydelse för utvärderingens mätegenskaper. Samtidigt finns från flera håll önskemål om att utvärderingarna skall inkludera uppgiftstyper som ger utrymme för betydligt mer komplexa prestationer.

Frågan kring ramverkets validitet inrymmer många aspekter, men fyra framstår som särskilt viktiga: (1) i vilken utsträckning har ramverket acceptans och legitimitet; (2) i vilken utsträckning ger ramverket signaler som utvecklar skolverksamheten i en önskvärd riktning; (3) i vilken utsträckning återspeglar ramverket undervisningens innehåll och former; och (4) vilken är ramverkets dimensionalitet?

Ramverkets acceptans

Som redan påpekats är ett ramverk alltid att betrakta som en politisk kompromiss, men det är naturligtvis även möjligt att identifiera innehållsliga/tekniska aspekter som skiljer ”bra” ramverk från ”mindre bra” ramverk. I huvudsak måste dock ett ramverks validitet bedömas med utgångspunkt från i vilken grad konsensus råder mellan de olika intressenterna; ett ramverk som förkastas av starka intressegrupper saknar validitet.

Detta betyder i sin tur att utformningen av den process med vilken ramverket genereras, och den utsträckning i vilken olika grupper har inflytande över processen är viktiga faktorer som påverkar ramverkets validitet. Vad gäller NAEP är det uppenbart att man varit framgångsrik när det gäller utformningen av ramverken: dessa synes möta stor och bred acceptans och även om utvecklingsarbetet innebär intensiva diskussioner och kontroverser mellan olika ståndpunkter har dessa diskussioner också ofta upplevts som konstruktiva, fördjupande och utvecklande.

NAGB är det organ som har ansvaret för att utveckla och fastställa ramverk för NAEP. NAGB har förvisso varit ett kontroversiellt organ, inte minst i samband med införandet

av rapportering i termer av prestationsnivåer. Icke desto mindre betraktas NAGB som ett organ med tillräcklig integritet för att dess beslut kring ramverken skall få acceptans. Instruktionerna till de uppdragstagare som genomfört arbetet med att specificera ramverken att göra det i former som tillförsäkrar en bred förankring har också varit tydliga.

För de internationella studierna är förankring av och konsensus kring ramverken inte lika tydlig, även om stora ansträngningar görs att involvera representanter för de olika länderna i utvecklingsarbete och beslut. I PISA är tanken att ramverket inte skall vara läroplansanknutet, utan snarare baserat på en föreställning om vilka kunskaper och färdigheter som är viktiga i det nuvarande och kommande samhället. IEA-undersökningarnas ramverk är i än större utsträckning än NAEPs ramverk kompromiss-produkter, som en följd av variationen i läroplaner mellan de deltagande länderna. Samtidigt kan det också noteras att ramverken för TIMSS och PIRLS uppvisar stora likheter med motsvarande ramverk för NAEP, liksom att de faktiska skillnaderna i ramverk och utfall för PISA och IEA-undersökningarna inte är så stor.

Ramverkets signalfunktioner

En betydande mängd forskning pekar på att prov och provuppgifter kan ha effekt på hur lärare och elever utformar verksamheten i skolan. Särskilt är detta tydligt när det gäller ”high-stake” prov, vilka ofta leder till en fokusering på det förväntade innehållet i provet, och på förberedelser i form av övning på det slags uppgifter som provet innehåller. Även ”low-stake” prov som NAEP-proven kan dock ha signalfunktioner, och redan i de första förslagen till utformning av proven betonades att uppgifterna inte endast skulle vara av flervalstyp, utan att elevproduktiva uppgifter skulle utgöra ett stort inslag. Även om andelen flervalsuppgifter blivit större än vad Tyler angav i sitt förslag till utformning av NAEP har under projektets gång ambitionsnivån successivt höjts när det gäller inslag av performance-upppgifter och uppgifter som kräver ”extended production”. Ett skäl för detta är givetvis att sådana uppgifter ses som mer informationsrika och tolkbara än flervalsuppgiter, men ett annat är också att NAEP av dem som utvecklar ramverken ses som ett effektivt instrument för att inspirera skolorna till nya grepp när det gäller metoder och tekniker för bedömning och utvärdering av elevers kunskaper och färdigheter.

Som redan påpekats är ett sådant synsätt i hög grad förenligt med Messicks utvidgade validitetsbegrepp, där ”consequential validity” bland annat står för de effekter som ett mätinstrument har på tänkande och på utformning av olika verksamheter. Den stora acceptans som NAEP har fått har säkert delvis sin grund i att de prov som används där i grunden skiljer sig från de individutvärderande prov som är så ymnigt förekommande i det amerikanska utbildningssystemet, och för vilka negativa effekter på inlärning och utveckling ofta noterats.

Det måste samtidigt konstateras att de uppgifter som bedöms ha positiva signaleffekter inte nödvändigtvis är de som är mest optimala vad gäller mätningarnas effektivitet. De produktiva uppgifterna är ofta mycket tidskrävande och mödosamma att besvara för eleverna, och det krävs stora insatser i form av tid och resurser för att bedöma elevernas produktion. Det finns också en tydlig tendens att dessa uppgifter har betydligt större

andel utelämnade svar än flervalsuppgifter. Det kan sålunda finnas en risk för en motsättning mellan ramverkets förmåga att ge lämpliga signaler, och dess förmåga att inom en rimlig kostnadsram ge en god grund för korrekta slutsatser om elevernas kunskaper och färdigheter.

Ramverket och elevernas undervisningserfarenhet

Som redan noterats är en vanlig kritik mot NAEP-uppgifterna och de uppgifter som ingår i de internationella undersökningarna att de tar upp innehåll som inte behandlats i undervisningen på en viss skola eller ett visst skolsystem. Det är dock uppenbart att idén med ramverk som använder sig av brett definierade kunskapsinnehåll har som konsekvens att vissa delar av det innehåll som finns i proven kommer att vara obehandlat i undervisningen för alla elever. Tanken är ju att ramverket skall fånga in den variation som finns och att resultaten skall vara tolkbara på aggregerad nivå. Det är därför fullt rimligt att förvänta sig att ett visst ramverk väl förmår att fånga ett innehållsområde på nationell nivå, men att det i mycket varierande grad svarar mot den undervisning som bedrivits på lägre aggregationsnivåer inom utbildningssystemet. Med Messicks (1989) terminologi kan vi säga att proven vidlåds av allt mer av ”construct underrepresentation” ju lägre nivå vi betraktar. NAEP-proven betraktas därför inte som användbara för att utvärdera undervisning på klass- och skolnivå, vilket var en anledning att diskussionen var intensiv innan beslut fattades om att resultat skulle presenteras för delstater och skoldistrikt.

Det är viktigt att inse att den form av mismatch mellan ramverkets innehållsliga specifikationer och den genomförda undervisningen som alltid föreligger i större eller mindre utsträckning då man använder sig av breda ramverk inte bara sätter gränser för de deskriptiva resultatens meningsfullhet på lokal nivå, utan att detta också allvarligt återverkar på möjligheten att förklara variationen i utfall med variabler som återspeglar olika former av resursinsatser. Om de prov som används för att mäta undervisnings-resultat endast delvis förmår att fånga upp resultaten är det inte heller möjligt att etablera samband med olika förklaringsvariabler.

Ramverkets dimensionalitet och innehåll

Ramverken är ofta organiserade som tvådimensionella system, där den ena dimensionen ofta avser olika innehållsdomäner (”content strands”) (i matematik exempelvis ”algebra och funktioner”, ”geometri”, ”taluppfattning”, ”sannolikhetslära och statistik”). Den andra dimensionen avser olika processer eller förmågor (exempelvis ”begreppslig förståelse”, ”proceduriell kunskap”, ”practical reasoning”). Därtill fogas ofta ytterligare indelningar. Trots detta fokuserar man i allmänhet i analys och rapportering på ett enda sammanfattande totalmått, även om man också ofta gör försök att komplettera redovisningen med resultat för olika underdimensioner definierade i innehållsliga och/eller processmässiga termer. En av anledningarna till den tydliga återhållsamheten i användningen av underdimensioner vid formuleringen av ramverken är att varje ytterligare dimension ställer krav på fler uppgifter för att det skall vara möjligt att nå acceptabel mätsäkerhet. En annan anledning är att erfarenheten ofta visat att underdimensionerna är mycket högt interkorrelerade, och att informationstillskottet av

analys och redovisning på underdimensionsnivå därför ofta är ringa (se t ex Zwick, 1987).

Mot bakgrund av diskussionen i föregående avsnitt av den stora betydelse som provens koppling till elevernas undervisningserfarenheter är det ett oväntat resultat att underdimensionernas betydelse är så liten i förhållande till en generell prestationsdimension, eftersom detta pekar på att provuppgifternas innehåll och utformning inte har så stor betydelse för resultaten. En möjlig förklaring till detta är att olika metodproblem medför att den generella faktorns betydelse överskattas i de dimensionalitetsanalyser som genomförts.

Ett sådant metodproblem har med matrissamplingsdesignen att göra. Denna medför nämligen att varje elev får ett mycket begränsat antal uppgifter som kan hänföras till en och samma underdimension, medan samtliga uppgifter influeras av den generella dimensionen. Eftersom de dimensionsanalytiska metoderna (framförallt olika former av faktoranalys) bygger på observerad samvariation mellan olika variabler medför detta problem att överhuvudtaget identifiera några underdimensioner, och att deras betydelse underskattas. En analys av läsproven i IEAs ”reading literacy” studie från 1991 (Elley, 1994), där inte matrissamplingdesign användes, tillsammans med läsproven från IEAs PIRLS-undersökning från 2001, i vilka matrissamplingdesign användes, visade på flera former av flerdimensionalitet i de förra proven, medan PIRLS-proven i större utsträckning var endimensionella (Gustafsson & Rosén, i tryck). Dessa resultat pekar på att matrissamplingen kan vara betydelsefull, men detta är ett område där ytterligare forskning behövs.

Ett annat metodproblem som kan förklara den generella faktorns dominans kan vara att resultaten påverkas av elevernas motivation, och andra faktorer som har genomslag i ungefär samma utsträckning på alla uppgifter, som exempelvis elevernas läs- och skrivförmåga. Sådana faktorer kan ses som metodfaktorer, vilka med Messicks (1989) termer bidrar med ”construct-irrelevant variance”.

En tredje tänkbar förklaring är att flerdimensionaliteten döljs av gruppskillnader. Muthén, Khoo och Goff (1994) anpassade flerdimensionella modeller till matematikdata från NAEP för olika undergrupper definierade efter kön och etnisk bakgrund, och kunde både visa på existensen av relativt starka underdimensioner, och att mönstret av gruppskillnader varierade över de olika dimensionerna. Trots dessa betydelsefulla resultat tycks inga fortsatta analyser med denna ansats ha genomförts på NAEP-data. Som tidigare nämnts förutsätter dimensionalitetsanalyser att det finns ett underlag i form av observerade kovarianser mellan olika uppgifter, och mönstret av kovarianser bestäms av hur matrissamplingsdesignen har utformats. Inom matematik används en BIB-design som i större utsträckning genererar kovarianser mellan uppgiftsblock än vad som är vanligt inom andra ämnesområden, vilket kan förklara den sparsamma förekomsten av dimensionalitetsanalyser av NAEP-data.

De resultat som presenterats ovan pekar på att det kan finnas anledning till en viss skepsis till slutsatsen att det inte finns anledning att uppmärksamma underdimensioner, därför att

de inte visats vara empiriskt betydelsefulla. Tvärtom framstår det som angeläget att ramverk och undersökningsdesigner utformas på ett sådant sätt att frågan om flerdimensionalitet kan få en adekvat belysning. Som visats av Kaplan (1995) kan konfirmatorisk faktoranalys som hanterar systematiskt bortfall med fördel användas i sådana analyser.

National Academy of Education har i sin serie utvärderingar ägnat ramverkens konstruktion betydande uppmärksamhet, och flera olika studier har utförts i vilka dessa granskats av olika kategorier av experter. Pellegrino et al. (1999) sammanfattar resultaten på följande sätt:

> … we conclude, on the basis of studies conducted previously, and on the committee's own observations, that NAEP's existing frameworks in science, mathematics, and reading generally reflect many goals of the disciplinary communities and have instituted some forward-looking, reform-oriented innovations. However, the frameworks still do not adequately reflect contemporary research and theory from cognitive science and the subject-area disciplines about how students understand and learn. Maintaining broad coverage of subject-area knowledge and skills is still a major focus of the frameworks, particularly in science and mathematics. Although breadth of coverage supports traditional assessment methodologies that result in summary scores as indicators of student achievement, it provides little insight about the level and depth of student understanding that is valued in many current views of student learning. (Pellegrino et al., 1999, s 128).

Även om man i dessa utvärderingar i princip ställer sig positiv till ramverken menar man sålunda att det finns ett utrymme för förbättring vad gäller en mer komplex och nyanserad beskrivning av arten och djupet av elevernas förståelse.

Uppgifternas validitet

Efter att ett ramverk konstruerats är nästa steg att konstruera ett lämpligt antal uppgifter i enlighet med ramverkets specifikationer. Det antal uppgifter som behöver konstrueras bestäms både av i vilken utsträckning som rapportering skall göras av olika delpoäng och av det antal uppgifter som behövs för att täcka av de olika innehållsområdena och uppgiftstyperna. Det totala antalet uppgifter och den totala testtiden bestämmer sedan, tillsammans med den mängd tid som finns tillgänglig för varje elev, hur många olika block och häften som behövs. Detta påverkar i sin tur hur många elever som behöver ingå i urvalet.

Konstruktionsarbetet innebär uttolkningar av de mer generella definitionerna och utsagorna i ramverket, med påföljd att en variation mellan olika konstruktörer är oundviklig. En annan fråga som är central i detta sammanhang är i vilken utsträckning de olika delarna av ramverket också får en tillräcklig representation bland de konstruerade uppgifterna. Ytterligare en fråga som är av stort intresse är i vilken utsträckning som

antalet uppgifter är tillfyllest för att ge en tillräckligt god täckning av ramverkets alla olika aspekter.

Pellegrino et al. (1999) redovisar resultat från flera undersökningar av i vilken utsträckning uppgifterna förmår att representera ramverkens intentioner. Huvudslutsatsen var att:

> … NAEP's assessments, as currently constructed and scored, do not adequately assess some of the most valued aspects of the frameworks, particularly with respect to assessing the more complex skills and levels and types of students' understanding (Pellegrino, et al., 1999, s 132).

Denna slutsats grundade författarna på flera olika observationer och resultat. Ett sådant var att andelen elever som väljer att hoppa över de mer krävande ”extended constructed-response” (ECR) uppgifterna är hög, vilket innebär att ett adekvat underlag för bedömning av de mer komplexa elevprestationerna saknas. Man noterar också att de bedömningsanvisningar som används för ECR-uppgifterna i allt för stor utsträckning är baserade på enkelt bedömbara aspekter av elevsvaren, men att variation i elevernas förståelsedjup inte väl fångades upp av anvisningarna. Det initiativ som togs vid formuleringen av ramverken för naturvetenskap i början av 1990-talet att varje elev skulle genomföra en laborativ uppgift såg man också som en källa till problem:

> Initially this appeared to be a laudable method for promoting hands-on learning experiences in science instruction. However, the evidence is mounting that such tasks, when administered in standardized fashion as part of a large-scale survey assessment, are not an adequate way to measure achievement in scientific investigations and related cognitive skills … The standardized tasks in the NAEP science assessment (and other large-scale survey assessments) are necessarily highly structured, have a very heavy reading load, and appear to measure some general reasoning skills and the ability to read and follow directions at least as much as the scientific investigation skills highlighted in the framework. Also, the generalizability of similar types of science performance tasks appears to be rather low. … The current technology for using performance-type measures in science (and in other NAEP subject areas) via the current large-scale survey assessment clearly has serious shortcomings. (Pellegrino et al., 1999, s 133).

Kommittén drog slutsatsen att det inte är möjligt att få en mer fördjupad och nyanserad bild av elevers tänkande och förståelse om man endast förlitar sig på de metoder som är tillgängliga i storskaliga undersökningar. De föreslog istället en multimetod ansats i vilken de storskaliga kartläggningarna kompletteras med mer intensiva undersökningsmetoder.

Slutsatser

Ett problem som uppmärksammats i flera utvärderingar av NAEP är att det medför svåra tolknings- och samordningsproblem att arbetet med ramverk, uppgiftskonstruktion, och utveckling av bedömningsanvisningar är uppdelat på olika grupper av personer och ansvariga institutioner. Forsyth et al. (1996) noterade:

> In the current NAEP scheme, subject area committees develop frameworks. Task developers attempt to create tasks that reflect the frameworks. Statistical analysts gather and analyze data that have the potential to determine the extent to which the tasks truly provide useful information about students' skills, as envisioned by the framework committees. But the feedback loop is never closed, forfeiting the opportunity to continually refine the subject-area specialists' vision with the actualities of the assessment. (Forsyth, et al., 1996, s 73).

Forsyth et al. (1996) föreslog därför att stående ämnespaneler bör inrättas. Även Pellegrino et al. (1999) noterade att utveckling av ramverk, uppgiftskontruktion, utprövning av uppgifter, sammanställning av den slutliga uppgiftsuppsättningen, och bedömning och analys inom NAEP ofta sker på ett alltför fragmenterat och okoordinerat sätt.

Denna genomgång av frågor kring ramverkens och uppgifternas validitet med fokus på NAEP pekar på att de uppfyller högt ställda krav vad gäller bredd och täckning utifrån traditionella ämnesformuleringar. Pellegrino et al. (1999) noterar dock att ramverken:

> … currently do not adequately capitalize on current research and theory about what it means to understand concepts and procedures, and they are not structured to capture critical differences in students' levels of understanding. They also do not adequately describe more comprehensive goals for student achievement that go beyond subject-matter knowledge and focus on the skills and abilities that will be important to the educated person in the next century.
>
> … advances in the study of cognition provide valuable insights into problem solving, explanation, interpretation, and how complex understanding is achieved, and they can be used to inform the development of assessments that better measure these dimensions of achievement than can the current array of broadly used large-scale assessment technologies. (Pellegrino et al., 1999, s 138).

Kommittén är också medveten om att en sådan höjning av ambitionsnivån inte kan genomföras inom ramen för de storskaliga undersökningsmetoder som nuvarande NAEP bygger på. Man föreslår därför en omstrukturering av nuvarande NAEP till ett multimetod NAEP:

> … a major component of this new paradigm NAEP is a core NAEP, consisting of large-scale survey assessments. Core NAEP would continue to track trends in achievement for both national NAEP and state NAEP in core subjects. Core subjects would include reading, mathematics, science and writing, and any other subjects, such as U. S. history or geography, in which assessments are

> administered frequently enough to establish trend lines. However, core NAEP alone cannot assess all important aspects of student achievement. The second major component in our proposed design is multiple-methods NAEP, consisting of alternative surveys and assessments. These components should be used to assess (1) components of core subject area frameworks that are not well suited for assessment via large-scale surveys, (2) nontrend subject areas, (3) achievements of members of special populations who cannot participate in the large-scale surveys, and (4) achievements of students with specific instructional experiences (e. g., fine arts, advanced mathematics).
> … Although we contend that a wider range of methodologies must have a place in a new paradigm NAEP to appropriately assess all aspects of the current frameworks and to be able to assess broader dimensions of achievement, we simultaneously recognize that this would simply not be feasible, financially or logistically, if it were assumed that all assessment methods were administered to a sample of students as large as those to whom the current large-scale survey assessments are administered. Smaller samples of students, and samples less fully representative of the nation should be used… (Pellegrino et al., 1999, s 147).

Förslag att NAEP skall delas upp dels i en "kärna" som är baserad på uppgifter som lämpar sig för användning i storskaliga undersökningssammanhang, dels i moduler som använder sig av andra uppgifter och metoder har framförts i flera andra sammanhang (t ex Forsyth et al., 1996). Det finns anledning att återkomma till dessa förslag längre fram.

Reliabilitet

I mätsammanhang syftar reliabilitetsbegreppet oftast på den grad av osäkerhet som vidlåder observation av individer. I NAEP och andra nationella utvärderingssystem är individresultat inte intressanta, utan dessa är endast hjälpmedel för att nå fram till skattningar av gruppkarakteristika. Kvaliteten på de observerade data från eleverna är icke desto mindre av avgörande betydelse för hållbarheten i de slutsatser som dras från utvärderingen.

Antal uppgifter per elev

Även reliabilitet i den traditionella bemärkelsen av grad av osäkerhet i de individuella observationerna är av stor betydelse i de grupporienterade utvärderingarna. Detta har sin grund i att det i dessa finns en direkt utbytbarhet mellan det antal uppgifter som varje elev besvarar, och det antal elever som måste ingå i undersökningen. Då varje elev besvarar få uppgifter, krävs i gengäld ett större antal elever i urvalet.

Det måste också betonas att även om ambitionen vid grupporienterad utvärdering är att undvika individuella värden, är detta inte möjligt för alla typer av frågeställningar. De frågor kring dimensionalitet som diskuteras ovan kräver exempelvis att det finns ett visst antal uppgifter för varje kombination av dimensioner vi önskar identifiera. Om utvärderingen har som syfte att bestämma grad av samvariation mellan kunskaper och färdigheter inom olika domäner går denna fråga inte att besvara genom att använda

matrissamplingsdesign inom varje domän, utan elever måste få uppgifter från mer än en domän. Frågeställningar som avser den grad av betydelse som olika determinanter har på kunskaper och färdigheter kräver också i vissa fall att vi har mätningar av god reliabilitet på individnivå.

En av anledningarna till att man inom NAEP valt att utforma designerna med få uppgifter per elev är att enskilda elevresultat enligt lag inte får registreras. En annan anledning är att deltagandet är frivilligt och en av de viktigaste faktorerna som bestämmer om en elev beslutar sig för att delta eller ej är den förväntade arbetsinsatsen. Ur många synpunkter skulle det dock vara fördelaktigt med flera uppgifter per elev. Kostnaderna blir lägre, vissa analyser förenklas, och mängden frågeställningar som går att belysa blir större.

Motivation

En av de viktigaste faktorerna som påverkar datakvaliteten är elevernas motivation. Eftersom NAEP-proven är ”low-stake”, med frivilligt deltagande, ingen rapportering av elevresultat, och ingen återföring av resultaten till eleverna kan elevmotivationen vara ett problem. Detta problem är särskilt markerat i åk 12, där deltagandefrekvensen är lägre än i åk 4 och 8, och där andelen överhoppade uppgifter är högre. Men även i de lägre åldrarna har problemen med deltagande och elevernas aktiva engagemang fått mycket uppmärksamhet. Inte minst har låg elevmotivation anförts som en förklaring till de låga resultaten i NAEP.

Vad gäller NEMP tycks elevmotivationen inte vara något problem, trots att varje elev medverkar under sammanlagt fem timmar utspridda under en vecka. En anledning till detta kan vara att NEMP endast inkluderar elever från åk 4 och åk 8, och en annan anledning kan vara att både innehåll och former för NEMP-uppgifterna framstår som mer varierade än för NAEP-uppgifterna.

”Low-stake” vs ”high-stake”

Linn och Baker (1996) sammanfattar resultatet från flera undersökningar som genomförts i syfte att undersöka i vilken grad motivationsfaktorer påverkar NAEP resultaten. I en serie undersökningar inkluderade man tidigare använda block av NAEP-uppgifter i utvärderingar på delstatsnivå som i större utsträckning var av ”high-stake” karaktär eftersom de avsåg ansvarsutkrävande på skolnivå. Resultaten visade på små skillnader och för endast två grupper av uppgifter förelåg statistiskt signifikanta skillnader, som även dessa var små. Dessa studier pekar på att motivationsförhållanden i NAEP inte behöver ha en negativ effekt på resultaten.

Linn och Baker (1996) redovisar också resultat från en serie experimentella studier, där man varierade faktorer som ekonomisk ersättning, tävlan och beröm, och jämförde med de vanliga förhållandena för administration av NAEP uppgifterna. Även i dessa studier var skillnaderna små, och endast i åk 8 fann man en effekt av ekonomisk ersättning (1 $ för varje korrekt besvarad uppgift). Effektstorleken var dock blygsam (0,20) och man drog slutsatsen att även dessa studier pekar på att ”… NAEP results do not seriously

understate student performance due to the low-stakes nature of the assessment" (Linn & Baker, 1996, s 15).

Uppgifternas art

Erfarenheterna inom NAEP visar dock också att elevernas benägenhet att besvara uppgifterna i hög grad påverkas av uppgifternas egenskaper, med höga svarsfrekvenser för flervalsuppgifter och låga svarsfrekvenser för uppgifter som kräver långa elevproducerade svar:

> A short constructed response task has the potential to tell us more about a student's thinking than a multiple-choice item. A longer constructed response task might tell even more about *some* students, but *nothing at all* about those who decide not to bother responding to it. Omit rates for grades 8 and 12 in 1994 Geography, for example, averaged less than 1 % for multiple-choice tasks and about 5 % for short constructed-response tasks – but up to 40 % for tasks in which students were asked to provide extensive responses. (Forsyth et al., 1996, s 16).

Det faktum att öppna och krävande uppgifter både ger ett stort svarsbortfall och även ofta ger svårtolkade utfall har lett till att man i flera utvärderingar har pekat på att det kanske skulle vara lämpligare att inte inkludera sådana uppgifter i "kärn" NAEP utan snarare låta dem ingå i specialstudier. Forsyth et al. (1996) diskuterar denna fråga med utgångspunkt i en distinktion mellan CR ("constructed response") och ECR ("extended constructed response") uppgifter:

> Both require open-ended responses, as opposed to MC items, but they differ notably in the amount of time and entry required of students. ECR tasks generally require more time, exhibit interconnections among aspects of performance, and can effect poor performance for a variety of reasons – only one of which is lack of requisite skills or knowledge. For example, Yepes-Bayara (1996) describes a talk-aloud study of sixteen 8^{th} grade students as they worked their way through both a "regular" block of science items from the 1996 assessment (i. e., MC and CR tasks) and a "special" block (i. e., either a ECR hands-on experiment or a "theme block" comprising several MC and CR tasks all dealing with the ecology of a pond). He found that the major difference between students who did well and poorly on the hands-on block was not a lack of science understanding, but trouble with planning and management skills (Forsyth et al., 1996, s 72-73).

Enligt Forsyth et al. pekar detta på att det inte är rätt sammanhang att lägga in ECR uppgifter i NAEP block och administrera dessa till tusentals studenter:

> … such tasks provoke considerable problems with nonresponse, low motivation, and poor performance for reasons other than the targeted skill. They take much more time than MC and CR tasks. The variety of reasons for poor performance often renders them uninformative from the technical perspective of their contributions to measurement accuracy, in terms of the domain of tasks in the

> subject area as a whole. They are most difficult for human scorers, and as such are fertile sources of scoring anomalies that must be identified and, if possible, rectified. Despite their fertile possibilities of providing more information about what students know and can do, they can result in a NAEP that actually tells less! Yet, to be as faithful as possible to the subject area frameworks, they are important to include in NAEP. (Forsyth et al., 1996, s 74).

Den lösning Forsyth et al. (1996) föreslår på detta problem är att endast inkludera MC och CR uppgifter i kärn-NAEP, och istället lägga ECR-uppgifterna i ett parallell-system. Detta förslag liknar i mycket det som lagts fram av Pellegrino et al. (1999).

Tidsomfattningen

En annan faktor som man erfarenhetsmässigt funnit vara av stor betydelse för elevernas deltagande är omfattningen av den insats som krävs av dem. Enligt Forsyth et. al (1996) har man observerat en kraftig negativ effekt på elevernas villighet att delta då den förväntade tidsinsatsen överstiger 50 – 60 minuter.

Begränsningar i den mängd tid som står till förfogande aktualiserar ett närliggande problem, nämligen förluster i datakvalitet som orsakas av en ambition att inkludera mer uppgifter i varje häfte än vad eleverna hinner med att besvara. Tanken är att NAEP proven inte skall vara snabbhetsbetonade, vilket innebär att ytterligare tid inte skall medföra högre resultat. Inom NAEP har man noterat att det ibland finns en tendens att uppgifter mot slutet av häftena besvaras i mindre utsträckning, vilket kan vara en indikation på tidsbrist. Man gör en distinktion mellan ”utelämnade” uppgifter och ”ej nådda” uppgifter, där de senare identifieras på så sätt att ingen av de efterföljande uppgifterna i häftet har besvarats. Linn och Baker (1996) noterade att:

> Nonresponse rates on items on the 1986 mathematics assessment were high enough to raise serious concerns that the results on that assessment might have been unduly influenced by speed. In the 1986 mathematics assessment, 23% (104 of 446) of the items had not-reached rates of .20 or higher. The not-reached rates were so high that the response rate criterion for including an item in the mathematics scale had to be relaxed so that items with not-reached rates as high as .45 were included in the scaling. (Linn & Baker, 1996, s 23).

Linn och Baker redovisar också andra studier av förekomsten av ej nådda uppgifter i olika NAEP undersökningar. Dessa visar att omfattningen av detta problem varierade mellan olika undersökningar, men också att det fanns skillnader mellan olika under-grupper i vilken utsträckning man hunnit besvarat uppgifterna, och dessa skillnader kunde endast delvis förklaras av skillnader i förmåga mellan grupperna. Med hänvisning till dessa resultat menar de att NAEP regelmässigt bör analysera och rapportera förekomsten av utelämnade svar.

Det kan noteras att IEAs ”reading literacy” undersökning som genomfördes år 1991 (Elley, 1994), och där inte matrissampling användes, var i hög grad påverkad av

snabbhetseffekter, vilka också hade stor betydelse för utfallet av länderjämförelserna (Gustafsson, 1997). En tanke med matrissampling är att den mängd uppgifter som varje elev skall göra hålls nere, och det är uppenbarligen ett framgångsrikt sätt att införskaffa prestationsdata på en stor mängd uppgifter. Samtidigt tycks det finnas frestelser att tänja på gränserna för antalet uppgifter i varje block, vilka det finns anledning att vara vaksam mot.

Slutsatser

Frågorna kring datakvalitet är i hög grad kopplade till validitetsfrågorna, och i många fall kan vi notera att det finns ett inverst samband mellan höga ambitioner vad gäller validitet och kvaliteten i de insamlade data. Den relativt omfattande användningen av ECR-uppgifter har sålunda medfört att det varit nödvändigt att minska uppgiftsantalet för varje deltagande elev. Många elever avstår också från att överhuvudtaget avge något svar på ECR-uppgifter. Ökning av antalet uppgifter inom häftena leder till att snabbhetseffekter uppstår, vilka i sin tur introducerar både slumpmässiga och systematiska fel. Utökning av antalet häften per elev leder till minskad motivation, med ty åtföljande bortfall av svar. Dessa erfarenheter indikerar att det är av yttersta vikt att det praktiska genomförandet av utvärderingen noga beaktas i samband med utveckling av uppgiftshäften och administrationsrutiner.

Mätning av förändring över tid

Bland de många syften som angivits för NAEP är det ett som av alla de utvärderingar som genomförts framhålls som unikt för NAEP och som särskilt viktigt, nämligen att studera förändringar i nivån av kunskaper och färdigheter över tid. Utvärderingarna framstår också som eniga i att NAEP varit framgångsrikt när det gäller att uppfylla detta syfte. Det är dock också uppenbart att trendmätning är förknippad med en lång rad metodologiska problem, och att man under årens lopp stött på en stor mängd problem i arbetet med att fastställa utvecklingen inom olika områden.

Som redan nämnts gör man trendmätning med två slags studier inom NAEP. Den ena är den undersökningsserie som benämns LTT (Long-Term Trends) och den andra görs inom ramen för huvud NAEP. Utvärderingarna av NAEP har noterat att det är både dyrt, komplicerat och förvirrande att trendmätning görs inom två studier som är helt separata, och som ibland ger motstridiga resultat. I utvärdering efter utvärdering har man har därför föreslagit att de två separata studierna skall integreras till en. Som framgår av den förteckning över planerade studier som visas i Bilaga 2 är dock tanken att LTT-studierna skall genomföras vart fjärde år inom matematik och läsning (men inte naturvetenskap) åtminstone fram till år 2016. Det faktum att det inte visat sig så lätt att integrera de två slagen av studier ger anledning att närmare granska de problem och utmaningar som finns i trendmätning.

Trendmätning inom huvud NAEP

Som redan påpekats har det varit en tanke med huvud NAEP att ramverk och uppgifter skall kunna ändras för att återspegla en aktuell syn på ämnesområdet, vilket i sin tur försvårar jämförelser över tid. Samtidigt är det ett uttalat syfte att även huvud NAEP skall ge underlag för studium av förändring över åtminstone kortare tidsperioder. Som noterades av Forsyth et al., (1996) skedde under perioden mellan 1984 och 1996 så många förändringar att man kunde konstatera att ”the only constant is change.” (s 33):

> Changes in frameworks, item specifications, the time of year of testing, age definitions, exclusion rules, and so on, have been the rule in NAEP. … Sometimes the impact of change is large enough to make it impossible to directly compare results from one assessment to the next. In all the years of NAEP assessments since 1984, it has only happened once that results from three successive assessments have been comparable (1994 Mathematics). Typically only two in a row are comparable before revisions sufficiently large to obviate the continuation of a ‘clean’ trend line are introduced. ‘One in a row’ is not uncommon. (Forsyth et al., 1996, s 34).

Efter 1996 har dock NAGB som en grundläggande policy uttalat att ramverk skall vara i huvudsak oförändrade under minst 10 år, för att göra det möjligt att skatta förändringar under minst denna tid. Det framstår också som om man nu nått fram till en betydligt högre grad av stabilitet i ramverk och procedurer i huvud NAEP, så att större tilltro nu kan fästas vid trendresultat från denna studie.

Inom huvud NAEP används metoden att med hjälp av IRT-teknik skapa skalor som är jämförbara över tid, men som bygger på att endast en del av uppgifterna är gemensamma. Detta innebär att en del av uppgifterna kan släppas fria efter användning.

Grundtanken är sålunda att de gemensamma uppgifterna gör det möjligt att erhålla mätningar som är jämförbara över tid. Detta kan dock genomföras på olika sätt. En möjlighet är att fixera uppgiftsparametrarnas värden till de värden som erhölls vid den initiala mätningen. En annan möjlighet är att skatta uppgiftsparametrarna för den nya mätningen separat, och sedan länka samman skalorna med en av de många tekniker som utvecklats för länkning via gemensamma uppgifter. En tredje möjlighet är att göra nya skattningar av samtliga uppgifters parametrar, varvid parametrarna för de uppgifter som är identiska över de olika mättillfällena skattas med samma värden vid samtliga tillfällen (flergrupps IRT). Inom NAEP har man i allmänhet använt en variant av den sist beskrivna ansatsen, genom att man successivt använt data från två på varandra följande mättillfällen.

Hedges och Vevea (2003) har presenterat resultat från en omfattande simuleringsstudie av olika ekvivaleringsmetoder med utgångspunkt från NAEP data. De fann att de olika metoderna överlag fungerade väl, med en i allmänhet försumbar bias. Den i särklass bäst fungerande metoden var dock den som baserades på flergrupps IRT:

> Multiple group IRT methods have considerable scientific method for equating and scale linking. These methods have the potential of practically eliminating bias in scale linking, even in the situations where current methods are weakest. (Hedges & Vevea, 2003, s 25)

Erfarenheterna från huvud NAEP av de IRT-baserade metoderna för att studera förändring är sålunda goda, så länge som ramverk och administrationsprocedurer är i huvudsak oförändrade. Sådana metoder har också använts inom TIMSS med goda resultat, och avsikten är att de skall användas även för trendstudier inom PIRLS. Tekniken för att studera förändring över tid med uppgiftsuppsättningar som är delvis olika tycks sålunda ha kommit över de initiala problemen, och nått god stabilitet.

En viktig bidragande orsak till dessa initiala problem var att den sanna förändring man sökte bestämma var mycket liten, eftersom populationsförändringar i allmänhet är mycket små över några få år. De procedurförändringar som infördes hade därför ofta större påverkan på resultatet än den ”sanna” förändring man försökte upptäcka. Det faktum att den förväntade förändringen är liten från ett mättillfälle till ett annat medför dock även andra metodologiska utmaningar. Som visas av Grissmer, Beaton & Hedges (2003) är ofta den trendmässiga förändringen mellan två på varandra följande mätningar mindre än ett medelfel, med påföljd att risken att dra felaktiga slutsatser om man endast jämför den aktuella mätningen med den närmast föregående mätningen på rent statistisk grund är stor. I synnerhet gäller detta naturligtvis i de fall där medelfelet är stort, på grund av ett litet urval personer eller ett begränsat antal uppgifter. Som visas av Grissmer et al., (2003) finns det därför anledning att i mån av möjlighet studera trendmässig förändring över längre tid med utnyttjande av all tillgänglig data.

Trendmätning inom LTT NAEP

Medan man inom huvud NAEP successivt byter ut uppgifter används inom LTT NAEP hela tiden samma uppgifter, vilka givetvis hålls hemliga. Anledningen till detta är som redan nämnts erfarenheterna från ”the 1986 reading anomaly”.

Även om man inte ändrar uppgifterna kan det dock invändas att mätinstrumentets egenskaper trots detta ändrar sig över tid. Ord blir exempelvis mer eller mindre vanliga över tid, vilket kan påverka texters svårighetsgrad. Som ett exempel kan nämnas att ordet ”separera” enligt resultat från UGU-projektet vid institutionen var betydligt svårare vid början av 1960-talet än 30 år senare. Förskjutningar av fokuseringen i matematik-undervisning över tid kan också påverka svårighetsgraden för olika kategorier av matematikuppgifter.

Man har dock, inte utan viss förvåning, kunnat konstatera att trendmätning med LTT-ansatsen fungerat väl:

> An interesting unintended positive effect has been the continuation of so-called ’long term trend’ assessments in Reading, Mathematics, and Science, which still

> use definitions, booklets, and administration procedures from the 1970's and early 1980's. Procedures for long-term trend have been refined and honed, so analyses needed for what would correspond to the 'standard report card' … can be carried out reliably and quickly – within 3 months after receipt of data. (Forsyth et al., 1996, s 34).

Man kan konstatera att LTT NAEP varit framgångsrikt när det gäller att beskriva trender som går tillbaka ända till år 1969 (se t ex Campbell, Hombo, & Mazzeo, 1999). Detta pekar på att stabiliteten vad gäller innehållsdomänerna och uppgifterna kan vara större än vad som det upplevda behovet att närmast kontinuerligt förändra och utveckla ramverken indikerar.

Zieleskiewicz (2000) har presenterat en intressant jämförelse mellan uppgifter i LTT-NAEP och huvud NAEP i vilken hon lät 30 lärare och ämnesexperter göra bedömningar av uppgifter i matematik och naturvetenskap för åk 8, dels från LTT-NAEP, dels från huvud NAEP. För varje uppgift besvarades först frågan: "Does this item assess knowledge and/or skills that students in your school will have covered in science (mathematics) by the end of eighth grade?" Om svaret på frågan var ja skulle också följande fråga besvaras: " … relative to all of what students cover in science (mathematics) in your school, how important is it for students to know/be able to do what is covered in the item by the end of the eighth grade". Svaret skulle avges på en skala från 1 till 5, med ändpunkterna "Not important" och "Very important". Två ytterligare frågor i vilka orden "will have covered" var utbytta mot orden "should have covered" skulle också besvaras.

Resultaten visade inte på några stora skillnader mellan frågor från de två NAEP-delarna. För frågan om innehållet behandlats i undervisningen fanns en svag tendens till högre medelvärden för LTT-NAEP (0,87) än för huvud NAEP (0,80), medan det inte var någon observerbar skillnad mellan uppgifter från de två undersökningarna vad gällde bedömningen av om innehållet borde behandlas (0,87 i båda fallen). Följande huvudslutsatser drogs:

- The long-term trend and main NAEP mathematics and science items evaluated in this study appear to reflect important content and skills that grade 8 teachers cover in their classes.
- This subset of items also appears to reflect what grade 8 teachers and disciplinary specialists believe are important parts of national standards in mathematics and science.
- For both mathematics and science there do not appear to be any meaningful differences in grade 8 teachers' or disciplinary specialists' perceptions of the relevance of long-term trend items and main NAEP items, as measured by coverage in current classrooms or reflections in national standards. (Zieleskiewicz, 2000, s 129).

Det faktum att bedömarna i denna undersökning inte förmådde differentiera mellan uppgifterna i de två NAEP-delarna ger en förklaring till varför LTT NAEP fungerat så väl, trots att uppgifterna i många fall är flera årtionden gamla.

Även om det med viss förvåning konstaterats att de i huvudsak oförändrade instrumenten inom LTT NAEP fungerat väl inom flera områden gäller detta inte generellt. Inom skrivning har sålunda LTT NAEP upphört från 1999 och inom naturvetenskap har den mätning som skulle genomförts 2004 skjutits på framtiden, för att det skall vara möjligt att utveckla nya uppgifter. Instrumenten har också omformats så att de bjuds i häften med roterande block på samma sätt som inom huvud NAEP. Den tidigare bandspelarstyrda administrationsmodellen har i samband med detta också övergivits. Stora förändringar har sålunda nyligen genomförts av LTT NAEP så att uppläggning och genomförande nu i hög grad sammanfaller med de som används inom huvud NAEP. Dessa förändringar har genomförts på så sätt att man genomfört ”bridging studies” i vilka både de nya och de gamla procedurerna använts på jämförbara urval av elever, för att det skall vara möjligt att ta hänsyn till eventuella effekter av de ändrade procedurerna.

Slutsats

Den grundfråga som undersökts här är: ”How can a single assessment be developed that is stable enough to provide long-term trends while still being flexible enough to adapt to changes in assessment approaches?” (Kolen, 2000, s. 133). Resultaten pekar på att detta kan åstadkommas både genom att inom ramen för väldefinierade och konstanta ramverk successivt byta ut uppgifter, och genom att använda samma uppgifter i samtliga mätningar.

Erfarenheterna visar dock också att resultaten är mycket känsliga för ändringar i de procedurer som används vid datainsamling varför det är angeläget att följa principen ”When measuring change, do not change the measure” oberoende av vilken av de två ansatserna man väljer.

En intressant fråga är givetvis i hur hög grad ramverk kan förändras utan att detta äventyrar jämförbarheten över tid. Eftersom resultaten visar på att utvärderingsresultat är i hög grad robusta över de olika underdimensionerna inom ett ramverk pekar detta på att möjligheterna att göra förändringar med bibehållen tolkbarhet i resultaten är goda. På denna punkt behövs dock fortsatt forskning innan det är möjligt att dra några bestämda slutsatser.

5. Fördjupad analys, tolkning och förklaring

Det råder stor enighet om att syftet att beskriva kunskaper och färdigheter är fundamentalt, och de utvärderingar som gjorts visar att både NAEP och andra nationella utvärderingssystem varit framgångsrika vad gäller att nå detta syfte. Oenigheten är större vad gäller övriga syften.

Ett syfte som många fäster stort avseende vid är tolkning och förklaring av de erhållna deskriptiva resultaten. Grundargumentet är därvid att beskrivningarna i sig är ointressanta om undersökningarna inte också bidrar med tolkningar och förklaringar som gör det möjligt att vidta åtgärder som leder till förbättringar. Mot detta argument står, som redan påpekats, ståndpunkten att även deskription kan initiera förändringar genom att resultaten påverkar aktörer på olika nivåer i systemet. Enligt detta synsätt är det systemets aktörer, och framförallt aktörer på den lokala nivån, som genom sina tolkningar och förklaringar vidtar åtgärder som är adekvata utifrån de lokala förhållandena.

Ett ytterligare argument mot högt ställda ambitioner vad gäller fördjupad analys, tolkning och förklaring är att det är praktiskt taget omöjligt att komma fram till korrekta slutsatser om kausala relationer inom surveystudier med tvärsnittsdesign, och många menar också att komplexiteten i de data som genereras av studier som använder den inom NAEP utvecklade metodologin är så stor att de inte är tillgängliga för fördjupad analys av den metodologiskt normalkunnige forskaren (se t ex Barron, 2000, Skolverket. 2004).

Dessa och andra argument diskuteras mer utförligt nedan.

Spontana resultattolkningar

Det kan vara lämpligt att först presentera resultat från studier av hur nationella utvärderingsresultat tas emot och tolkas av olika grupper, och i anlutning till NAEP har flera undersökningar gjorts av detta (t ex Hambleton & Meara, 2000; Hambleton & Slater, 1996).

Pellegrino et al. (1999, s 25-26) redovisar resultat från undersökningar av hur NAEPs beskrivande resultat tagits upp och presenterats av media, och i vilken utsträckning resultaten tolkats korrekt av olika grupper av användare. Bland annat har man studerat hur tidningar rapporterat resultaten från NAEP 1990, 1992 och 1994, liksom av den första delstats-NAEP. Man fann att tidningarna ägnade NAEP-resultaten stort intresse, och att detta i synnerhet gällde de delstater där resultaten var mindre bra. Man observerade också att presentation och diskussion ofta gick långt utöver vad data och analys egentligen tillät, och att man gjorde mycket fria tolkningar av orsakerna till de svaga resultaten. Ett tydligt mönster var att såväl politiker som nyhetsmedia gärna knöt an till populära föreställningar och gängse politiska käpphästar när det gällde att förklara alternativt bortförklara de svaga resultaten. Ofta innebar detta att man hänförde förklaringar till variabler och faktorer som inte ens fanns med i NAEPs design, ”… in

other cases, interpretations went beyond the data to bolster commentators' preconceived notions or already established political agendas." (Pellegrino et al., 1999, s 26).

Man noterade också att kommentarerna ofta innehöll långtgående och konkreta förslag till hur resultaten skulle kunna förbättras, t ex genom förbättrad lärarutbildning, mer utbildningsresurser, och tidigare diagnos av elevers inlärningsproblem. Återigen saknade dock dessa förslag koppling till NAEPs resultat.

Slutsatsen av analysen var att det finns ett stort och legitimt behov att identifiera orsakerna till både goda och dåliga resultat, och att försöka använda dessa förklaringar till att förbättra utbildningen.

Pellegrino et al. (1999, s 27-35) redovisar också resultat från en egen undersökning av hur resultaten från 1996 års NAEP av matematik och naturvetenskap presenterats och tolkats i pressen. I undersökningen klassificerade man excerpter från ledande dags-tidningar i tre kategorier: deskriptiva, värderande och tolkande. De deskriptiva resultaten, som ofta formulerades i jämförande termer, var i allmänhet korrekt återgivna i den meningen att de stämde med det som rapporterats av NAEP. När det gällde värdering och tolkning av resultaten gick man ofta långt bortom data och de resultat som kunde grundas på NAEPs design. Man säger:

> In our view, the excerpts … demonstrate how users hope and try to use NAEP results to inform thinking about the performance of the education system, schools and student groups. As was observed for earlier administrations, some NAEP users accorded more meaning to the data than was warranted in laying out reasons for strong and weak performance. Others sought to better understand strengths and weaknesses in students' knowledge and skills. (Pellegrino et al., 1999, s 31).

Hambleton och Meara (2000) rapporterar resultat från en än mer omfattande analys av tidningars sätt att redovisa resultat från NAEP-undersökningarna, och de kommer fram till liknande slutsatser. De noterade också en tendens att man gjorde synnerligen fria tolkningar av orsakerna till de deskriptiva resultaten, men också att det fanns en tydlig tendens att samband tolkades i kausala termer, utan att man var medveten om att detta inte är möjligt. Negativa samband mellan omfattningen av TV-tittande och resultat på NAEP-uppgifter tolkades exempelvis så att det var det myckna TV-tittandet som orsakade de dåliga kunskaperna och färdigheterna. Författarna drog slutsatsen:

> There is substantial evidence to suggest that the NAEP press release materials are being used and understood by the press. The problems occur when press tries to go beyond the materials provided and interpret the findings for the public. A careful reading of the NAEP releases leads us to feel that NCES and NAGB may not want to interpret the complex array of findings for the public: they may want the readers to arrive at their own interpretations. However, this goal may be unrealistic because (1) quantitative literacy is not high across the country … and (2) the data and NAEP reports are themselves very complex. (Hambleton & Meara, 2000, s 150).

Denna starka tendens hos politiker och media att gå bortom det tillgängliga empiriska underlaget i försöken att hitta förklaringar till de deskriptiva resultatmönstren är givetvis beklaglig, eftersom det innebär ett slags missbruk av utvärderingsresultaten. Men det finns knappast anledning att förfasa sig över att dessa kategorier av avnämare ägnar sig åt fria tolkningar av orsakerna till utvecklingen inom skolområdet. Samma agerande, om dock ibland i något mer subtila former, ägnar sig forskare åt när de försöker finna förklaringar till varför kunskapsutvecklingen i skolan gestaltar sig på det ena eller det andra sättet. Även forskare tenderar att föra fram den egna favoritteorin som den primära förklaringskällan, oavsett om det finns en hållbar empirisk grund för detta eller ej. Samma agerande är för övrigt lätt att återfinna inom den statliga skoladministrationen, där den handlingsplan (”Resultatförbättring i grundskolan”) som Skolverket publicerade 2005-05-02 är ett aktuellt exempel på hur tolkningar och förklaringar till den negativa utvecklingen av elevernas kunskaper och färdigheter läggs fram till trots av en mycket svag kunskapsmässig underbyggnad.

Det tycks sålunda vara ett faktum att ett nationellt utvärderingssystem inte kan begränsas till att ha ett renodlat deskriptivt syfte, eftersom de deskriptiva resultaten i sig kräver att man kommer fram till tolkningar och underlag för att förbättra skolan. Behovet av att också kunna lägga fram väl underbyggda tolkningar och förklaringar av de observerade resultatmönstren har många gånger identifierats även inom NAEP. NAGP har sålunda konstaterat att ”… current presentations of NAEP results by grade, demographic group, state and a small number of additional variables do not point policy makers and educators to possible sources of disappointing or promising performance or to their possible policy implications.” (Pellegrino et al., 1999, s 31).

Kausala slutsatser från tvärsnittsstudier

Ett grundläggande problem är att det ofta är omöjligt att dra slutsatser om orsaker till goda eller dåliga resultat på grundval av resultat från tvärsnittsstudier, därför att de metodologiska utmaningar som måste övervinnas för att dra kausalt korrekta slutsatser är för stora för att de skall kunna klaras av inom ramen för studier med sådan design.

Exempelvis visar flera NAEP-studier av läsning i åk 4 (t ex år 1992, Mullis et al., 1993, 126-127) på ett negativt samband mellan mängden direkt läsundervisning elever fått och deras läskompetens. Det är emellertid knappast rimligt att detta samband kan tolkas i kausala termer, på så sätt att en större mängd läsundervisning skulle orsaka en sämre läskompetens. En omvänd kausalitet, där elever som är svaga i läsning i åk 4 är utsatta för en mer omfattande direkt läsundervisning, framstår som en mer rimlig tolkning av det negativa sambandet. Det metodologiska grundproblemet här är givetvis att det föreligger en selektionsbias, i det att de elever som fått olika mängd läsundervisning inte hade samma initiala läskompetens. Selektionsproblemet är ofrånkomligt i tvärsnittsstudier av det slag som NAEP representerar, och utgör ett utomordentligt allvarligt hot mot möjligheterna att dra korrekta slutsatser om effekten av undervisnings- och resurs-variabler. Anledningen till detta är att undervisningsinsatser praktiskt taget aldrig fördelas

slumpmässigt över olika elever, utan med hänsynstagande till olika former av bedömda behov. Ofta leder detta till olika former av kompensatorisk resursallokering, (dvs att elever med större behov får mer och/eller bättre undervisning), vilket tenderar att skapa ett observerat negativt samband mellan undervisning och resultat (Gustafsson & Myrberg, 2002).

Det faktum att de enkla bivariata sambanden mellan undervisningsvariabler och inlärningsresultat praktiskt taget aldrig kan tolkas i kausala termer gör att de ofta snarare skapar förvirring än bidrar till ökad insikt och förståelse. Detta är dock den huvudsakliga formen i vilka analysresultat från NAEP rapporteras, och det har ifrågasatts om det är ändamålsenligt att presentera resultat som så lätt kan misstolkas (Pellegrino et al., 1999).

Problemet med selektionsbias kan helt undvikas endast då man använder experimentell design, med slumpmässig fördelning av försökspersoner över olika betingelser. Experimentell metodik är dock knappast aktuell i samband med nationell utvärdering. Longitudinell design, som ger möjlighet till analys av ’value-added’, ger jämfört med tvärsnittsdesign bättre möjligheter att dra korrekta kausala slutsatser, även om inte heller sådana studier är utan sina problem. I samband med NAEP har det dock aldrig varit aktuellt att överväga en longitudinell design, bland annat eftersom denna förutsätter att enskilda individer kan identifieras och följas över tid.

Även med tvärsnittsdesign kan det dock ibland, under förutsättning att adekvata variabler har mätts, vara möjligt att med hjälp av statistisk metodik kontrollera för selektionsbias. Med exempelvis regressionsanalys eller strukturell ekvationsmodellering kan man i statistisk mening likställa grupper, och på så sätt dra korrekta kausala slutsatser. Ett problem är dock att den information som måste finnas tillgänglig för att fullt ut åstadkomma en sådan statistisk kontroll sällan är tillgänglig. I detta sammanhang är mått på tidigare prestation praktiskt taget oundgängliga, och i NAEP finns inte sådana. Istället har man utnyttjat indikatorer på elevens socioekonomiska bakgrund som kontrollvariabler, vilka i vissa fall är adekvata, men i andra fall är dåliga substitut.

Ett annat problem är att den typ av komplex statistisk analys som är nödvändig för att dra kausala slutsatser från tvärsnittsdata är utomordentligt svår att genomföra på komplexa data av det slag som NAEP bygger på. Här används komplexa stratifierade klusterurval vilka förutsätter att analysalgoritmerna kan ta hänsyn till den påverkan detta har på punkt- och felskattningar. Vad som är ett än större problem är att matrissamplingstekniken leder till stora svårigheter då data skall analyseras med statistiska tekniker där samtidig hänsyn tas till många variabler. Detta gör det till ett komplext och mödosamt företag att analysera NAEP-data med en högre ambitionsnivå (Barron, 2000).

I sin utvärdering intar Pellegrino et al. (1999) ståndpunkten att även om det är svårt att nå fram till bestämda tolkningar och entydiga slutsatser om kausala relationer från utvärderingsdata är det angeläget att ansträngningar görs att så långt det är möjligt tolka och förklara utvärderingsresultaten.

> … education policy based on imperfect data is better than education policy with no empirical base. We believe that the benefits of documenting interrelationships among achievement and educational variables in ways that respect the complexity of the educational enterprise will outweigh its disadvantages. (Pellegrino et al., 1999, s 39).

I den analytiska modell Pellegrino et al. skisserar är dock inte avsikten att hela det empiriska underlaget skall samlas in inom ramen för NAEP med hjälp av enkäter till elever, lärare och skolledare. Istället är avsikten att data från andra undersökningar skall samordnas med prestationsdata från NAEP för att därigenom skapa ett mer heltäckande och komplett indikatorsystem:

> We seek a system that suggests relationships among student, school and achievement variables and that stimulates democratic discussion and debate about American education. We believe that … housed in broader system of education indicators, NAEP results can help drive and increase NAEP's relevance to policy research. Like Bohrnstedt (1997:10), we believe that a coordinated system will "provide a very fertile basis for hypothesis generation and the preliminary generation and the preliminary exploration of ideas about what works and doesn't in American education." It is our position that providing associative data about issues of concern to educators, policy makers, and the public will prompt more deliberate exploration and explanation of the interrelationships among achievement and educational variables. It is our hope that the system's products will be used to pose hypotheses about student achievement and test them, moving beyond observational to experimental research and using longitudinal designs. (Pellegrino et al., 1999, s 39).

Trots att man tänker sig ett relativt ambitiöst system kan det noteras att man ändå är relativt återhållsam när det gäller möjligheterna att dra bestämda slutsatser. Man ser det snarare som en källa till frågor och hypoteser, som sedan måste undersökas med andra metoder i andra sammanhang.

Fördjupad analys

Diskussionen ovan har framförallt fokuserat på möjligheten att nå fram till en mer fördjupad förståelse av vilka faktorer som påverkar utvecklingen av kunskaper och färdigheter. Det måste dock också påpekas att det inte går att göra en skarp gränsdragning mellan beskrivning å ena sidan, och förklaring å den andra sidan. Anledningen till detta är att beskrivningsmodellerna måste vara baserade på teoretiska föreställningar om vad som skall observeras, hur aggregeringar till sammanfattande mått skall göras, och hur redovisning och analys skall göras. Ett exempel kan närmare illustrera det nödvändiga samspelet mellan beskrivning, fördjupad analys och tolkning.

Under perioden från 1992 till 2003 var resultatet i läsning för elever i åk 4 praktiskt taget oförändrat: för populationen som helhet hade resultat ökat med 1 poäng från 217 till 218 poäng, vilket inte var statistiskt signifikant. Men en närmare analys av data visade

samtidigt att olika etniska undergrupper i materialet i allmänhet förbättrat sina resultat: vita med 8 poäng från 224 till 229 poäng; svarta med 6 poäng från 192 till 198 poäng; spanska med 3 poäng från 197 till 200 poäng; och asiatiska med 10 poäng från 216 till 226 poäng. Vid första anblick förefaller det orimligt att populationen som helhet kan ha ett oförändrat resultat när resultaten för samtliga undergrupper utvecklats i positiv riktning! Förklaringen till paradoxen ligger i att det skett andelsmässiga förskjutningar mellan undergrupperna mellan 1992 och 2003. Den relativt högpresterande gruppen av vita minskade sin andel av populationen från 73 % till 60 % under denna tid, medan den relativt lågt presterande undergruppen med spanskspråkig bakgrund ökade från 10 % år 1992 till 17 % år 2003. Detta är ett exempel på vad som brukar kallas Simpsons paradox, vilken innebär att förändringar i en variabel kan dölja förändringar relaterade till en annan variabel om de två variablerna är korrelerade.

Detta exempel illustrerar att det kan vara nödvändigt att även i den grundläggande deskriptiva rapporteringen av utvärderingsresultat ansätta analytiska modeller som tar hänsyn till exempelvis demografiska förändringar. Detta är dock inte utan sina problem, vilket illustreras av den diskussion som förts kring användning av olika tekniker för att justera för demografiska förändringar i rapporteringen av NAEP-resultat. Finn (2004) har argumenterat mot användning av sådana justeringstekniker:

> … at a time when our premier education goal is to close race-related achievement gaps, it is bizarre to settle for academic outcomes adjusted for "demographics and family economics." Such statements imply that poor and minority kids ought not be expected to attain proficiency. (Finn, 2004)

Wainer (1994) har i en analys av NAEP-resultaten i matematik för år 1992 för New Jersey använt sig av tekniken att justera den demografiska strukturen för alla delstater så att den sammanfaller med den som gäller för hela USA. Denna justering medförde ett betydligt bättre resultat för New Jersey jämfört med andra delstater. Sådana analyser har emellertid förbjudits av NAGB, eftersom det kan betraktas som ett slags ”fusk” att göra kontrafaktiska justeringar.

I NAGBs invändning ligger givetvis att det inte är självklart vilka aspekter som skall tas hänsyn till då man justerar för demografiska och strukturella faktorer. I det exempel som beskrevs ovan gjordes en justering av befolkningens demografiska sammansättning med avseende på etnicitet, men vad är det som säger att inte även andra faktorer ska tas hänsyn till? Det är exempelvis rimligt att anta att den andel av föräldrarna som har postgymnasial utbildning har ökat mellan 1992 och 2003 och eftersom det finns ett samband mellan föräldrarnas utbildningsnivå och elevernas läsprestation borde detta innebära bättre resultat vid det senare mättillfället, om allt annat är konstant. Det torde i själva verket finnas ett stort antal variabler som hänsyn borde tas till om man skall göra justeringar, vilket innebär att mer grundläggande principer måste etableras för vilka variabler som skall användas, och hur justeringar skall genomföras.

Detta exempel illustrerar dock att det finns ett utomordentligt stort behov av fördjupad analys av data från nationella utvärderingsstudier även då dessa inte i första hand syftar

till att fastställa kausala relationer mellan variabler. Genom att föra in ytterligare variabler som exempelvis gör det möjligt att analysera resultaten för olika undergrupper, eller konstanthålla förändringar i andra variabler, kan de deskriptiva resultaten på aggregerad nivå bli mer meningsfulla och en mer detaljerad deskription kan erhållas. Även om det inte är möjligt att komma fram till entydiga slutsatser om kausala relationer gör detta det utomordentligt angeläget att förutsättningarna för fördjupade analyser är så goda som möjligt. Inte minst är det viktigt att adekvata bakgrundsvariabler har observerats, och att det är möjligt att i efterhand komplettera med ytterligare information om exempelvis skolors egenskaper och resurser.

Data från såväl NAEP som de internationella studierna har också använts för ett utomordentligt stort antal sekundäranalyser av forskare inom olika discipliner. Samtidigt är det mycket vanligt att man konstaterat att endast en bråkdel av den tillgängliga informationen utnyttjats för sekundäranalyser. För att stimulera forskare att utnyttja NAEP data för sekundäranalyser har NCES utlyst forskningsanslag och stipendier, och man har också kurser och workshops för forskare som är intresserade av att använda sig av NAEP data i sin forskning.

En av anledningarna till att man tvingas tillgripa denna form av åtgärder är att de data som samlas in med NAEP metodologin är utomordentligt svåra att analysera på ett korrekt sätt. Det kan för övrigt nämnas att dessa svårigheter inte endast upplevs av forskare med svaga metodologiska kunskaper. Braun (2005), som är framstående statistiker och som under många år lett forskningen vid ETS, har redovisat personliga vedermödor i hanteringen av NAEP data inom ett stort reanalysprojekt. Den stora källan till problem var här avsaknaden av reliabel information på individnivå.

Som redan nämnts utgör den komplexa NAEP metodologin en stor tillgång över vilka de involverade känner stolthet (Beaton & Johnson, 2004), men samtidigt utgör den en källa till problem och då framförallt i samband med reanalyser. Det finns anledning att ta upp dessa problem till en något mer utförlig diskussion.

Analysernas komplexitet

NAEP metodologin är komplex på flera olika sätt och det finns skäl att diskutera de olika problem som detta medför, och hur de kan hanteras.

Urvalsmodeller

I NAEP används flerstegs klusterurval för att välja ut de elever som skall ingå i utvärderingen (Rust & Johnson, 1992) och i allmänhet används likartade urvalsmodeller i de internationella studierna. För de nationella samplen användes under 1980- och 1990-talen en urvalsmodell där man först väljer ut geografiska områden, antingen på så sätt att vissa stora områden är garanterade att ingå i urvalet, eller på så sätt att sannolikheten att ingå i urvalet bestäms av antalet elever i området (”probability proportional to size”, PPS). I nästa steg gör man ett urval av skolor från vart och ett av dessa områden, varvid PPS-principen också används. I det tredje och sista steget gör man ett slumpmässigt urval

av elever inom de utvalda skolorna. För de urval som ingår i övriga NAEP-komponenter används varianter på denna flerstegsdesign.

Denna urvalsmodell innebär att olika individer har olika sannolikhet att komma med i urvalet. Eftersom sannolikheterna kan beräknas är det dock möjligt att korrigera för de olika urvalssannolikheterna genom att tilldela individerna vikter, som är omvänt proportionella mot sannolikheten att ingå i urvalet. Dessa vikter måste alltid appliceras i beräkningarna eftersom annars resultaten inte blir korrekta. Detta innebär en viss ökning av komplexiteten i bearbetningarna, även om modern statistisk programvara i allmänhet erbjuder enkla metoder för att analysera viktade data.

Vad som emellertid är ett allvarligare problem är att viktning av data medför att formelapparaten för att beräkna skattningarnas grad av osäkerhet (t ex i form av medelfel och konfidensintervall) blir betydligt mer komplex än då individerna ingår med lika vikt i urvalet. Inom NAEP har man löst detta problem genom att använda olika former av återsamplingstekniker (t ex ”jackknife” eller ”balanced repeated replications”, see Johnson & Rust, 1992). Dessa metoder ger korrekta resultat, men de är komplexa att använda, och de medför begränsningar i vilka analytiska metoder som är tillgängliga. Under senare år har dock statistisk programvara utvecklats som på analytisk väg ger skattningar av estimatens medelfel (t ex AM, Mplus 3, LISREL 8.72).

Som framgår av beskrivningen ovan är urvalen klusterurval, i det att hela skolor, och ibland även hela klasser, utgör urvalsenheter. Klusterurval medför, liksom viktade urval, implikationer för bestämning av skattningarnas grad av osäkerhet och det är viktigt att ta hänsyn till detta, inte minst därför att osäkerhetsgraden i ett klusterurval ofta är betydligt högre än i ett individurval av motsvarande storlek. Detta har att göra med att de elever som finns vid samma skola eller i samma klass tenderar att vara mer lika varandra än ett slumpmässigt urval av elever, vilket medför en informationsförlust vid klusterurval. Även då intraklasskorrelationen, som är ett mått på individernas grad av likhet inom klustren, är relativt måttlig, blir ofta medelfelen minst dubbelt så stora vid klusterurval som vid individurval, vilket har att göra med att relativt många individer ingår i de valda klustren.

Av de individer som väljs ut att ingå i en undersökning är det dock inte alla som faktiskt deltar, på grund av att skolor eller elever avstår från att delta, eller på grund av frånvaro från skolan vid det aktuella tillfället. Under senare år har man i NAEP märkt en tendens till ökning av olika former av bortfall och i synnerhet i åk 12, vilket kan ha att göra både med NAEPs ”low-stake” karaktär, och med en ökad omfattning av olika former av testningar i de amerikanska skolorna.

Ett sätt att hantera bortfall på skolnivå, är att använda ersättningsskolor, av vilket erfarenheterna synes vara goda. Ett annat sätt att hantera bortfall av både skolor och individer är att justera vikterna, på så sätt att de kvarvarande eleverna inom en skola ges högre vikt.

En annan typ av bortfall är svarsvägran på enskilda frågor, och inom NAEP har man noterat att andelen som besvarar en uppgift i stor utsträckning påverkas av uppgiftens

egenskaper. Uppgifter som kräver långa elevproducerade svar får sålunda en låg andel svar, vilket också är fallet för uppgifter av öppen karaktär med oklara uppgiftskrav. Bortfall av svar på enskilda frågor är svår att hantera.

Ytterligare en betydelsefall fråga i samband med urvalet gäller vilka elever som ingår i urvalsramen, och vilka som inte ingår i denna. Under det senaste årtiondet har denna fråga varit central i samband med delstats-NAEP, eftersom olika principer för att definiera populationen har kommit att hota möjligheterna till korrekta jämförelser. Det har varit möjligt för skolledningen på de deltagande skolorna att utesluta elever med fysiska eller mentala funktionshinder, eller med mycket begränsade kunskaper i engelska, från deltagande. Under slutet av 1990-talet har emellertid NAEP vidtagit olika åtgärder för att minska uteslutning av elever, både därför att syftet är att undersöka vad *alla* elever kan, och för att öka jämförbarheten. Detta har bland annat inneburit ökade möjligheter att anpassa proven och genomförandet av proven så att elever med olika former av funktionshinder kan delta. Olika försök har också gjorts att på statistisk väg erhålla skattningar av vilka resultat som skulle erhållas om det inte förekom uteslutning av elever (McLaughlin, 2001).

Matrissampling och plausibla värden

Som redan nämnts introducerades MML-skattning i NAEP-metodologin i samband med att ETS tog över NAEP år 1983, och med denna skattningsteknik är det möjligt att göra direktskattning av parametrar på gruppnivå utan att några individvärden skattas. För att göra det möjligt att distribuera data för sekundäranalys infördes tekniken att dra 5 slumpmässigt valda värden (”plausibla värden”) från den förväntade fördelningen av individuella förmågevärden, givet individens svar på uppgifterna och individens värden på ett antal bakgrundsvariabler. Om de plausibla värdena sedan hanteras korrekt är det möjligt att återskapa de estimat som erhålls genom direktskattning, och givetvis också göra ytterligare analyser. Detta kan genomföras med speciell programvara (t ex AM, eller med makron skrivna för SAS och SPSS, vilka distribueras tillsammans med data). Dessa program vidlåds dock av vissa begränsningar i utbudet av tillgängliga analystekniker. Senare versioner av HLM-programmet för flernivåanalys (Raudenbush, Bryk, Cheong & Congdon, 2004) hanterar automatiskt plausibla värden, vilket även gäller Mplus (Muthén & Muthén, 2004).

I avsaknad av speciell programvara för att hantera plausibla värden rekommenderas följande procedur: (1) genomför den önskade analysen 5 gånger, en gång för varje plausibelt värde; (2) bestäm standardavvikelsen i de fem estimaten för den eller de parametrar som är i fokus för intresset; och (3) denna standardavvikelse uttrycker den grad av osäkerhet som har sin grund i samplingen av uppgifter, och kan adderas till osäkerhetsskattningar som avser urval av individer. Detta är en relativt enkel procedur, om dock något omständlig, och det är denna som implementerats i HLM och Mplus.

Den föreslagna proceduren är givetvis korrekt, men jag vill för egen del uttrycka tvivel om detta alltid är det optimala sättet att hantera de plausibla värdena. Variationen i dessa utrycker ju grad av osäkerhet i mätningen av de enskilda individerna, eller

reliabilitetsbristerna, och sådan variation medför systematisk bias i skattningen av vissa parametrar (t ex regressionskoefficienter). I sådana fall förefaller det mig naturligare att beräkna ett medelvärde av de plausibla värden, alternativt använda dem som indikatorer på en latent variabel i strukturella ekvationsmodeller.

Med den inom NAEP använda tekniken verkar det överhuvudtaget inte vara möjligt att få ett enda optimalt individvärde. Den teknik som används inom PISA skapar dock dels 5 plausibla värden på samma sätt som inom NAEP, dels ett värde som är den bästa möjliga skattningen av individens förmåga.

Ett ytterligare sätt att komma tillrätta med problemen att hantera plausibla värden inom NAEP är att utnyttja de speciella procedurer för datananalys (bl a NAEP Data Tool) som finns tillgängliga via webben. Dessa beskrivs närmare nedan.

Betingade estimat

Det som kanske dock är den komplikation som ställer till mest bekymmer både i produktion av NAEP data och i analys av data är det faktum att bakgrundsvariabler används i analysen för att öka precisionen i skattningarna. Som redan nämnts bestäms fördelningen av individens förväntade förmågevärden dels av individens svar på uppgifterna, dels av ett antal bakgrundsvariabler (t ex föräldrarnas utbildning, skola, osv). Eftersom sådana bakgrundsvariabler lätt kan förklara 50 % av den individuella variationen i prestation i USA medför denna typ av information betydande tillskott av information, och i synnerhet i de fall då varje elev endast besvarar ett begränsat antal uppgifter är detta mycket värdefullt.

Det har ifrågasatts om det är etiskt korrekt att en individs resultat inte endast beror av prestation, utan även av olika variabler som individen normalt inte har någon kontroll över. Denna kritik måste dock ses mot bakgrund av att de individuella plausibla värdena inte skall betraktas som skattningar som skall användas och tolkas på individuell nivå, utan endast som hjälpmedel för att komma fram till skattningar av de olika populationskarakteristika.

En annan, och allvarligare, kritik är att problem kan uppstå då man använder sig av bakgrundsvariabler både som hjälpmedel för att öka precisionen i skattningarna och som analysvariabler. Det mest uppmärksammade problemet är att en bias kan uppstå i skattningar av effekter av individvariabler som inte funnits med som betingningsvariabler, men som används som analysvariabler. En lösning på detta problem är givetvis att se till att inkludera samtliga variabler som betingningsvariabler.

Barron (2000) pekar på ytterligare ett allvarligt problem:

> Another problem that conditioning causes for analysis is more fundamental. Researchers with years of experience with NAEP and strong backgrounds in statistics said that they still do not understand the methodology used to scale NAEP in anything but more than general terms and are unsure of the impact the

> scaling procedures have on analyses they have conducted or wish to conduct. They widely reported being uncomfortable using data in their research when they do not understand the scaling methodology used to generate the data. (Barron, 2000, s. 183).

Inom PISA används inte betingning i direktskattningarna, men liksom inom NAEP används betingning då de plausibla värdena beräknas för de enskilda individerna.

NAGB gav i direktiven för den redesign av NAEP som påbörjades vid mitten av 1990-talet uttryck för önskemål om att tekniken att använda information om bakgrundsvariabler vid skattning av elevresultat borde överges, men detta har av effektivitetsskäl inte gjorts. Den används också fortfarande inom TIMSS och PIRLS.

Nya analysverktyg

Som redan nämnts kan analyser av NAEP data numera genomföras med special-utvecklade analysverktyg som är tillgängliga via webben. *NAEP Data on the Web* är ett lättanvänt program för att plocka fram information ur alla genomförda utvärderingar från 1990, och som inte kräver detaljkunskap om undersökningar, variabler eller analys-tekniker. Med verktygets hjälp är det lätt att konstruera tabeller och grafik för jämförelse mellan olika grupper och undergrupper, och tabellerna går också lätt att flytta in i generella program som EXCEL.

Med *NAEP Data Tool Kit* går det också att göra korstabeller och även mer avancerade analyser som regressionsanalyser. Medan *NAEP Data on the Web* vänder sig till en bred krets av användare, riktar sig *NAEP Data Tool Kit* mer till forskare. Speciella tillstånd krävs dock för att komma åt data med detta verktyg.

Som redan nämnts har ett speciellt program utvecklats (AM; Cohen, 2002) för att genomföra MML-estimeringar och för att generera och analysera plausibla värden. Programmet erbjuder också en rad olika statistiska analysmetoder i vilka plausibla värden hanteras med automatik.

Även om dessa nya analysverktyg förvisso underlättar sekundäranalyser av data från NAEP och andra projekt som använder NAEP metodologi innebär även dessa begränsningar, och ställer även de stora krav på att användaren förstår datas egenskaper.

Sammanfattningsvis måste slutsatsen dras att det verkar svårt att komma bort från den stora komplexitet som introducerades i NAEP vid övergången till IRT-teknik under 1980-talets början. Utvecklingen av nya statistiska analysmetoder och webbaserade hjälpmedel gör det lättare att hantera komplexiteten, men metodologin ställer ändå betydande krav på resurser och kompetens.

Bakgrundsvariabler

Med termen ”bakgrundsvariabler” betecknas inom de nationella utvärderingsprojekten i stort sett samtliga variabler, utom de uppgifter som används för att mäta kunskaper och färdigheter. Sådana variabler kan tjäna tre olika huvudfunktioner, och ofta används en viss variabel för alla dessa ändamål:

1. Som betingningsvariabler för att förbättra precisionen i mätningen av kunskaper och färdigheter;
2. Som oberoende variabler i fördjupade analyser; och
3. Som föremål för analys och beskrivning.

Det har i många utvärderingssammanhang pekats på att många av de mest intressanta resultaten från NAEP avsett den bild av skola och undervisning som framträtt genom beskrivningar av olika bakgrundsvariabler.

Inom NAEP har graden av uppmärksamhet på bakgrundsvariabler och omfattningen av sådana varierat över tid. Under den första generationens NAEP var det antal bakgrunds-variabler som insamlades för förklaringsändamål mycket begränsat, vilket hade sin grund i att Tyler var tveksam till möjligheterna att utvinna någon kunskap genom att relatera bakgrundsvariabler till resultat. I samband med att projektet flyttades till ETS var tanken att ambitionsnivån skulle höjas vad gäller förklaringsorienterade analyser med hjälp av de mer kraftfulla mätmetoderna och med kraftigt utökat antal bakgrundsvariabler. De enorma mängderna data som samlats in blev dock aldrig föremål för ordentlig analys, vilket delvis hade sin grund i att ETS fick koncentrera ansträngningarna på att utveckla de nya skattningsteknikerna för mätningarna av kunskaper och färdigheter. Ambitionsnivåerna har därefter sänkts vad gäller systematisk insamling och analys av bakgrundsvariabler och de har framförallt kommit till användning som betingnings-variabler. I den senaste lagstiftningen kring NAEP lyfts dock forskningssyften fram på ett mer explicit sätt än tidigare.

Många forskare har uttryckt tvivel kring användbarheten hos bakgrundsdata i NAEP, och har bland annat pekat på att de enskilda frågorna ofta har så låg reliabilitet att de inte kan användas i analyser. Det är intressant att notera att denna typ av kritik ofta framförs av forskare som är mycket väl medvetna om vilka ansträngningar som måste göras för att med hjälp av många olika uppgifter åstadkomma reliabla och valida mätningar av kunskaper och färdigheter. En enskild fråga i ett frågeformulär är dock i allmänhet varken mer eller mindre reliabel än en enskild fråga i ett kunskapsprov, så det är svårt att förstå varför det finns förväntningar att det skulle vara lätt att mäta bakgrundsvariabler med frågeformulär.

Det är lätt att konstatera att utvecklingen av frågeformulär har fått en mycket styvmoderlig behandling i NAEP och de internationella studierna. Barton (2002) konstaterade angående NAEP:

> If school, teacher and student questionnaires administered to a large national sample were a free standing survey research operation, a considerable budget

> would be allocated for their development and maintenance. A considerable budget would also be allocated for data manipulation, analysis and reporting.
>
> There is, of course, a very sizeable budget for the development and analysis of the cognitive side, as there needs to be… The background question side has never been adequately funded as the large scale research effort that it has every appearance of being. (Barton, 2002, s 26).

Enligt Barton, som arbetat vid ETS Policy Information Center under lång tid med sekundäranalys av NAEP data, finns det både anledning och möjlighet att förstärka mätningen av bakgrundsvariabler i NAEP. Framförallt pekar Barton på det stora värdet av att kunna göra beskrivningar av resultaten för undergrupper som bestämts utifrån fler variabler samtidigt, och där det också finns möjlighet att kontrollera för inflytande av exempelvis föräldrarnas socioekonomiska status och hemmiljön. Han utesluter inte heller analyser av relationer mellan variabler i syfte att finna förklaringar, men menar att detta i första hand är analysarbete som bör utföras av forskare vid universiteten.

Barton pekar också på att det är angeläget att lämna analys av enskilda frågor till förmån för sammansatta mått i form av index och andra former av sammansatta variabler. Yang-Hansen, Rosén, & Gustafsson (i tryck) har visat att faktorpoäng beräknade från latenta variabelmodeller kan vara en framkomlig väg när det gäller att skapa sådana sammansatta variabler.

Det är uppenbart att bakgrundsvariabler är av stor betydelse i utformningen av nationella utvärderingssystem, liksom att det är en viktig uppgift att utforma instrument som mäter sådana variabler på ett sätt som gör det möjligt att nå utvärderingens syften.

6. Rapportering och värdering av resultaten

Som beskrivits tidigare var den första generationen av NAEP fokuserad på enskilda uppgifter och grupper av uppgifter. Även om denna beskrivningsnivå visade sig vara opraktisk vid jämförelser mellan grupper och för beskrivning av kunskapsutveckling över tid hade den dock den stora fördelen att ge konkret och förhållandevis enkelt tolkbar information. De IRT-baserade skalor som efterföljande generationer av NAEP har gått över till är höggradigt abstraherade och för att de skall kunna ges en mer innehållslig tolkning behöver de kunna knytas till prestationer på olika uppgifter. En lång rad olika tekniker har utvecklats vilka har som syfte att bibringa ökad tolkbarhet till IRT-skalorna.

Utformningen av rapporteringen är i hög grad avgörande för i vilken utsträckning ett nationellt utvärderingsprojekts syften skall nås, och det är också uppenbart att en av de centrala tvistefrågorna är hur resultaten skall beskrivas, presenteras och bedömas. Det kan finnas skäl att därför diskutera dessa frågor något mer ingående.

Innehållsnära beskrivningsmodeller

De första rapporterna från NAEP redovisade resultaten per uppgift i form av p-värden, och med separatredovisning för ett fåtal jämförelsegrupper (t ex kön, och de fyra regionerna). Längre fram övergick man till att redovisa resultaten för kategorier av uppgifter.

Under förutsättning att uppgifterna är offentliga ger detta en mycket informativ beskrivning, och då i synnerhet för dem som sysslar med undervisningsnära verksamhet och läroplanarbete. När man använder samma uppgift över flera tillfällen ger denna ansats också god information om förändring över tid, även om detta givetvis förutsätter att uppgifterna hålls hemliga under den tidsperiod de används.

Inom NEMP används i huvudsak den ursprungliga NAEP modellen för rapportering:

> So that results can be understood in relation to what the students were asked to do, national monitoring assessments are reported task by task. The results given alongside each task tell the relative percentages of students who demonstrated skills or knowledge to the standards prescribed by the marking criteria. About two-thirds of the tasks attempted by students in any given year are released this way. (Flockton, 1999, s 67).

Huvudargumentet för att välja denna model är att man i första hand önskar nå ut med resultaten till verksamheten, för att de skall kunna utnyttjas för att förbättra undervisningen:

> The prime goal of national monitoring is to contribute to the quality and improvement of student learning. Essential to achieving that goal is the availability of information which can be used to make judgements, support

> practice and guide decisions. That information needs to be accessible to the widest possible community of interest if it is to influence discussion and debate.

Man kan uppfatta en tydlig skillnad mellan den verksamhets- och innehållsnära beskrivning som NEMP fokuserar på å ena sedan, och den abstraherade beskrivning som NAEP sedan 1983 gått över till. Det måste dock understrykas att denna skillnad är delvis skenbar. Sedan uppgifter och resultat inom NAEP kunde göras tillgängliga via webben genom verktygen *NAEP Data on the Web* och *NAEP Questions Tool* har detta kommit att bli en mycket viktig del av NAEPs rapportering. Lazer (2004) skriver:

> This system represents a major revolution in data dissemination. Thousands of people visit the web tool each month, and state NAEP coordinators have been trained in its use. In addition, a web-based tool for disseminating released test questions is available and has become the single most visited portion of the NAEP website. Future versions of the tool will have an enhanced graphics capacity and will include the capacity to create tables and results not included in the current database of summarized results (Lazer, 2004, s 486).

Denna interaktiva form av analys och presentation har av Carr (2004) beskrivits som den tredje fasen i NAEPs tekniska utveckling.

Även Pellegrino et al. (1999) betonar att det finns ett starkt behov av innehållsnära analys och beskrivning av NAEP-resultaten, men menar också att det finns behov av betydligt mer sofistikerade beskrivningsmodeller:

> In contrast to the approach currently employed in NAEP, contemporary cognitive theorists would argue that inferences about the nature of a student's level of knowledge and achievement in a given domain should not focus on the individual, disaggregated bits and pieces of information as evidenced by questions that students can answer correctly. More important is the overall pattern of responses that students generate across a set of items or tasks. The pattern of responses reflects the connectedness of the knowledge structure that underlies the conceptual understanding and skill in a domain of academic performance. (Pellegrino et al., 1999, s 139).

Författarna för en utförlig diskussion om vilket observationsunderlag som skulle behövas för att det skall vara möjligt att komma fram till sådana mer fördjupade beskrivningar, och menar att det måste bygga på en kombination av olika metoder, där de storskaliga surveyteknikerna kompletteras med mer intensiva metoder som riktas till mindre grupper av elever. De menar dock också att även den information som samlas med den nuvarande metoduppsättningen är användbar för fördjupad analys och beskrivning, och framhåller som förebildliga exempel det arbete som utförts av "the National Council of Teachers of Mathematics" (NCTM) som har skrivit:

> … interpretive reports based on the analysis of students' responses to individual NAEP items. These reports … characterize student performance at different levels

> of detail appropriate for different audiences. For example, the most recent monograph, reporting on the sixth mathematics assessment, administered in 1992, includes an analysis of students' understanding of basic number concepts and properties, their computational skills, and their ability to apply number concepts and skills to solving problems, based on examinations of items that assess these skills and concepts … The report includes data across approximately 100 individual NAEP items. Patterns of responses are analyzed to draw conclusions about student performance on specific topics …
>
> The NCTM interpretive teams have consistently documented that the most critical deficiency in students' learning of mathematics at all ages is their ability to apply the skills that they have learned to solve problems. … The analyses also provide a perspective on relations between skill acquisition and the development of understanding of fundamental concepts. These conclusions, based on interpretive analyses of students' responses, address issues that are at the core of public debate regarding curriculum choices. (Pellegrino et al., 1999)

Sådana fördjupade analyser föreslår Pellegrino et al. (1999) bör genomföras inom alla ämnesområden med viss regelbundenhet.

IRT-baserade beskrivningsmodeller

Som tidigare beskrivits introducerades IRT-tekniken i NAEP 1983 därför att den tidigare matrissamplingsmodellen och beskrivningar på uppgiftsnivå inte tillät analys av fördelningar, och inte var praktiskt användbar i analys av undergruppers resultat. Trots en del initiala problem och en hög komplexitetsgrad får IRT-modellen beskrivas som utomordentligt framgångsrik, och de analyser och beskrivningar som den använts till har varit mycket användbara. Problemet är dock att de siffervärden som presenteras inte har någon omedelbar innebörd, och i synnerhet inte för personer som saknar erfarenhet av statistisk analysmetodik. En stor utmaning som ägnats mycket tankeverksamhet och uppfinningsrikedom är att finna sätt att ge en innebörd åt siffervärdena, och flera olika metoder har utarbetats.

En teknik som har använts inom NAEP är så kallade uppgiftskartor (”item maps”). En uppgiftskarta visar en IRT-skala från de lägsta poängen till de högsta poängen, och bredvid denna skala visas beskrivningar av de uppgifter som elever med olika poäng kan förväntas besvara korrekt. En uppgiftskarta har sålunda som syfte att beskriva vilken typ av kunnande som elever på olika nivåer på skalan uppvisar. Även om uppgiftskartor har stort värde vidlåds de också av begränsningar. Ett problem är att endast ett relativt begränsat antal uppgifter kan redovisas på de olika nivåerna, dels av utrymmesskäl, dels av det skälet att antalet uppgifter ofta är begränsat på grund av behovet att hemlighålla uppgifter för framtida bruk. Ett annat problem är att en sannolikhetsgräns för korrekt svar på uppgiften måste ges, och en sådan gräns är alltid subjektiv. Inom NAEP har man använt sannolikhetsgränsen 65 % för produktiva uppgifter, och 74 % för flervals-uppgifter, men dessa måste betraktas som godtyckliga.

En annan metod som använts inom NAEP är den så kallade skalankringsmetoden (Beaton & Allen, 1992). I ett första steg väljer man ut specifika punkter på skalan, så kallade ankringspunkter. Därefter identifierar man de uppgifter som i stor utsträckning besvarats korrekt av elever med resultat i närheten av ankringspunkten, och som också i stor utsträckning besvarats inkorrekt av elever med resultat vid den närmast lägre ankringspunkten. Slutligen genomför en kommitté av ämnesexperter en granskning och beskrivning av de kunskaper och färdigheter som kännetecknar elevprestationer vid de olika ankringspunkterna. Denna teknik kan dock också kritiseras för att den har inslag av subjektivitet, och resultaten är inte heller enkla att kommunicera.

Standardsbaserad rapportering

Under 1980-talet kom i ökande utsträckning krav på att utbildningsresultat skulle bedömas mot så kallade ”standards”. Standardsbaserad rapportering av resultat har också i allt större utsträckning kommit att användas inom NAEP. Innebörden i ”the standards movement” är att man skall fastställa vad som är godtagbara kunskaper på olika nivåer. Man talar dels om ”content standards”, som är beskrivningar av vad eleverna förväntas lära sig, dels om ”performance standards” vilket är olika sätt på vilka eleverna kan visa att de nått målen. Som påpekats av Skolverket (2003, s 150) är den svenska motsvarigheten till content standards ”mål att uppnå”, medan performance standards motsvaras av betygskriterierna. Skolverket påpekar dock också att de svenska målformuleringarna är hållna i mycket allmänna ordalag, medan de utländska motsvarigheterna i allmänhet är betydligt mer preciserade.

Tankarna om att NAEP skulle rapportera resultaten i termer av prestationsnivåer kom först upp i den tidigare nämnda Alexander-James kommisionen, som föreslog att en av uppgifterna för NAGB skulle vara ”… identifying feasible achievement goals for each of the age and grade levels to be tested.” (cit från Vinovskis, 1998, s 41). Den panel som kommenterade Alexander-James rapporten gick längre i att formulera krav på rapportering i termer av prestationsnivåer:

> For each content area NAEP should articulate clear descriptions of performance levels, descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent and expert. Descriptions of this kind would be extremely useful to educators, parents, legislators, and an informed public. (cit från Vinovskis, 1998, s 42).

Det främsta skälet för att begära denna typ av resultatbeskrivning var att man inte såg de skalor med innehållsbefriade siffervärden som genererades av de statistiska procedurerna som särskilt informativa eller meningsfulla.

> What does a level 400 on a reading test mean? Such scores can be used for comparison across time and localities, but the nation’s report card would be more broadly informative if it provided clear descriptions of the levels of competence demonstrated by our children. Much more important than scale scores is the reporting of the proportions of individuals in various categories of mastery at

> specific ages. In several fields, particularly reading and mathematics, we are in a position to describe beginning, average, and advanced competence at various ages. In other areas, such as writing, science, and computer literacy, research remains to be done. (Alexander & James, 1987, cit från Vinovskis, 1998, s 42).

Det första steget i arbetet var att precisera innebörden i termen ”appropriate achievement”, som var den som användes i beslutstexten, där man föreslog följande:

> In its goal-setting plan NAGB intends to base its definition of ”appropriate achievement goals” on knowledge and skills a consensus of educators and others say is needed to achieve the next level of subject-matter mastery. For 12th grade the Board intends to expand this consensus-building process to include employers and members of the public, college professors and scholars, to define the knowledge and skills all students need to participate in our competitive economy. We also propose to define the levels of proficiency needed to handle college-level work. (NAGB Briefing Book, 1990, cit från Vinovskis, 1998).

Processen började på lägre åldersnivåer, och först nyligen har man på allvar har påbörjat arbete med att etablera standards för åk 12 genom samverkan med universitet, military och närinsliv.

Tidigare försök på 1980-talet att etablera standards hade praktiskt taget enbart fokuserat på att etablera en lägsta ”godkänd”-nivå (minimum competency). För de behov NAGB identifierade för NAEP var detta dock inte tillräckligt, utan man formulerade sig i termer av tre nivåer. Den allmänna definitionen för dessa var:

> *Proficient*. This central level represents solid academic performance for each grade level tested – 4, 8, and 12. It will reflect a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12 the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.
> *Advanced*. This higher level signifies superior performance beyond grade-level mastery at grades 4, 8 and 12. For the 12th grade the advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams.
> *Basic*. This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade – 4, 8 and 12. For 12th grade this will be higher than minimum competency skills (which normally are taugtht in elementary and junior high schools) and will cover significant elements of standard high school-level work. (National Assessment Governing Board, 1991, cit från Vinovskis, 1998, s 45).

Det första försöket att koppla samman de abstrakta definitionerna av prestationsnivåer med NAEP-uppgifterna ägde rum år 1990. En stor grupp av 63 bedömare (huvudsakligen lärare) använde då en variant av Angoff-metoden för att etablera performance standards. Denna procedur innebär, i sammanfattning, att bedömargruppen gör en skattning av hur stor sannolikheten är att en elev på en viss prestationsnivå svarar rätt på en viss uppgift. Statistiska sammanställningar av dessa bedömningar utgör sedan grund för att fastställa de poänggränser på provet som skiljer de olika grupperna. Det visade sig dock vara en långt ifrån enkel uppgift för bedömargruppen att göra skattningarna, och stora inkonsistenser både mellan och inom bedömare uppenbarade sig. Efter tre bedömnings-omgångar nådde man dock fram till ett förslag som NAGB kunde anamma.

Denna process och dess resulterande produkt fick utstå mycket kritik. Inte minst var en utvärderingsrapport av Stufflebeam, Jaeger och Scriven (1991), vilka följde processen noga, utomordentligt kritisk. Detta väckte ont blod hos NAGB, och det sätt på vilket de försökte tysta ner kritiken och göra sig av med de misshagliga forskarna fick en hel del uppmärksamhet i massmedia (se Vinovskis, 1988, s 46-50). Kritiken mot det sätt på vilket NAGB fastställt gränserna för de olika prestationsnivåer ledde också till att General Accounting Office (GAO, dvs den amerikanska Riksrevisionen) utvärderade arbetet. I en rapport år 1993 med titeln *Educational achievement standards: NAGB's approach yields misleading interpretations* dömde man ut NAGBs arbete som "procedurally flawed" och av "doubtful valdidity". Framförallt ifrågasatte man den tekniska kompetensen hos NAGB och menade att denna styrelses uppdrag framförallt var att formulera policy, och att man inte borde blanda samman detta med operativt och tekniskt komplicerat arbete.

Även National Academy of Education (NAE) utvärderade standards i 1990 och 1992 års utvärderingar av läsning och matematik och kritiserade de använda procedurerna (Glaser et al., 1993). Liksom andra kritiker pekade man på stora skillnader mellan olika bedömare i panelen både av de olika uppgifterna och av var poänggränserna borde gå. Man menade också att de fastställda poänggränserna var för höga. En senare utvärdering, som också den initierades av National Academy of Education, riktade också mycket skarp kritik mot NAEPs procedurer för att etablera standards (Pellegrino, Jones & Mitchell, 1999), och betecknade dessa som "fundamentally flawed".

Inte minst som en följd av den massiva kritiken vidtog NAGB ändringar i de använda procedurerna, och lämnade också över till ACT att från 1992 genomföra bestämningen av standards. Man genomförde också konferenser där forskare diskuterade olika tekniska lösningar, med särskilt fokus på Angoff-tekniken. Även om inte någon enighet nåddes, var diskussionen i detta sammanhang mer nyanserad.

Brown (2000) har gått igen samtliga de utvärderingar som genomförts under 1990-talet av NAGBs procedurer för bestämning av standards, och diskuterat de olika kritik-punkterna. Han påpekar:

> … at the beginning of the 90's, NAGB was clearly troubled by reporting on the average score of the Nation as the referent of quality. NAGB believed that qualitative reporting would provide an impetus for change even if the performance levels of students were not satisfactory initially. The policy decision to establish performance levels proved to be as technically complex, as it was forward thinking. During the past decade, NAGB has persisted with its policy decision, even in the face of considerable criticisms from noted psychometricians who labeled the achievement level-setting process as flawed. (Brown, 2000, s 38)

Brown beskriver sålunda NAGB som förutseende, men också naiva angående de tekniska utmaningar som detta projekt innebar. Samtidigt menar han att det skett en kraftig förändring och förbättring av de procedurer som används för att fastställa standards:

> The results of the research conducted by ACT have improved the standard-setting model considerably. Changes in the panelists' training have been noteable. The additional sources of input for the various rounds of ratings and the use of participant feedback on the process have made today's standard-setting process much more refined and satisfactory than was the case at the beginning of the process. Much of the criticism in the literature, however, is of the process used in the first and second standard-setting sessions. Clearly the problems that were identified with the initial process were concerns to be addressed, and many of them have been studied. (Brown, 2000, s 38).

Även om det fortfarande pågår en intensiv diskussion och forskning kring olika metoder för att fastställa standards verkar det som om de av NAEP använda metoderna successivt kommit att få allt större acceptans, och sedan 1990 har standardsbaserad rapportering använts som ett komplement till de övriga rapporteringstyperna. Reckase (2000) har också gjort en inventering och genomgång av egenskaper hos olika metoder för att fastställa standards, och hans slutsats är att det ännu inte finns någon metod som är lika välundersökt och välfungerande som den modifierade Angoff-metoden. Även Hambleton et al. (2000) har ifrågasatt mycket av kritiken som oberättigad.

Slutsatser

Frågan om val av beskrivningsmodell är uppenbarligen nära kopplad till de syften som är formulerade för utvärderingen, och utformningen av beskrivningsmodellen är också i hög grad påverkad av vilka typer av uppgifter som används. Det är dock svårt att frigöra sig från intrycket att det är nödvändigt att använda sig av flera olika beskrivningsmodeller samtidigt. Utan sammanfattande, abstraherade, mått framstår det som omöjligt att göra beskrivningar av förändringar över tid, eller göra fördjupade analyser av exempelvis skillnader i resultat mellan olika grupper av individer. Det är också en stor fördel om resultaten av dessa analyser kan redovisas på sätt som är mer intuitivt tillgängliga. Utan de mer innehållsorienterade uppgiftsnära beskrivningarna framstår det också som svårt att mer direkt dra nytta av resultaten i utveckling av praktisk undervisningsverksamhet.

7. Överväganden och förslag avseende utformning av ett nationellt kunskapsbedömningssystem

Nedan diskuteras utformningen av ett övergripande ramverk för ett nationellt kunskapsbedömningssystem för den svenska grundskolan. Först diskuteras de syften ett sådant system bör ha, och överväganden görs kring systemets grundläggande design. Därefter diskuteras den mer detaljerade utformningen av kunskaps- och trendmätningar, liksom utformningen av informationsunderlag för förklaringsinriktade studier.

Tidigare har termen ”nationell utvärdering” använts, men i samband med diskussionen kring förslaget kommer istället beteckningen ”nationell kunskapsbedömning” att användas. Skälet till detta är att termen ”utvärdering” avser värdering av grad av måluppfyllelse, vilket exempelvis var det uttalade syftet med NU-03. För att undvika sådana associationer är det lämpligt att inte använda termen ”utvärdering” i detta sammanhang. ”Nationell kunskapsbedömning” är också en mer direkt svensk motsvarighet till engelskans ”national assessment” än uttrycket ”nationell utvärdering”.

Syften

Av uppdragsbeskrivningen (se kapitel 1) följer att det primära syftet skall vara att ge beskrivningar av kunskapsnivåer och av utvecklingen av nivå och spridning i kunskaper över tid på nationell nivå, dels totalt, dels för olika undergrupper (t ex kön, social bakgrund, utländsk bakgrund).

Systemets främsta syfte skall sålunda vara att ge information om utvecklingen över tid av kunskaper och färdigheter, varvid både nivå och spridning skall uppmärksammas.

Ett andra syfte med det nationella kunskapsbedömningssystemet är att ge underlag för analyser av orsaker till förändringar i kunskapsnivåer.

Även om det finns begränsningar i möjligheterna att nå fram till entydiga förklaringar till de observerade trenderna skall det nationella kunskapsbedömningssystemet i görligaste mån skapa underlag för analyser med sådana syften. Detta görs i första hand genom att information samlas in för potentiellt relevanta förklaringsvariabler.

Genomgången av de olika nationella utvärderingssystemen visar att dessa fyller en lång rad syften, där de beskrivande och förklarande syftena i allmänhet är de primära. Ofta betonas också syftet att systemet skall bidra till förbättring och utveckling av utbildning och undervisning på olika systemnivåer. Detta kan ses som det allra mest överordnade

syftet, som de två ovan formulerade syftena kan föras tillbaka på, men som i en så allmän formulering inte bidrar med vägledning.

För vissa system betonas att de skall bidra till lärares kompetensutveckling vad gäller bedömning och värdering av elevers resultat. Det framstår dock inte som lämpligt att göra detta till ett framträdande syfte för ett svenskt nationellt kunskapsbedömningssystem, eftersom kompetensutveckling av lärare i första hand är en uppgift som åvilar andra aktörer som Myndigheten för Skolutveckling, lärarutbildningen och kommunerna. Resultat och metoder framtagna inom den nationella kunskapsbedömningen kommer dock givetvis att vara betydelsefulla underlag i detta arbete

På grundval av observationer att provsystem har styreffekter på undervisning betonas också ofta att nationella kunskapsbedömningssystem genom att ge goda exempel på utformning och bedömning av uppgifter kan ha utvecklande effekter. Detta är en viktig funktion, som dock bättre uppfylls av det nationella provsystemet, dels därför att det i detta ingår ”high-stake” prov som görs av nästan samtliga elever, dels därför att uppgifter i dessa prov är offentliga efter genomförandet. Det nationella provsystemet omfattar också en provbank som kan ges vidgade uppgifter.

Även om de utvecklande och förebildliga funktionerna inte bör utgöra huvudsyften för den nationella kunskapsbedömningen är det givetvis viktigt att denna genomförs med högt ställda kvalitetskrav vad gäller val och utformning av uppgifter och bedömningsanvisningar.

Grundläggande designfrågor

Nedan diskuteras grundläggande frågor kring utformningen av det nationella kunskapsbedömningssystemet.

Ämnen och ämneskombinationer

Enligt direktiven skall samtliga grundskolans ämnen omfattas av den nationella kunskapsbedömningen. Uppdraget bör också tolkas så att inte endast de lätt utvärderingsbara ämnesdelarna skall fokuseras.

En av grundtankarna i den ursprungliga designen av NAEP var att hela skolans ämnesbredd skulle omfattas. Det har dock det skett en successiv förskjutning mot några få ämnen, och då framförallt mot läsning och matematik. Det finns flera anledningar till detta: de kan betraktas som viktigare än andra ämnen eftersom de är grundläggande för inlärning även inom andra ämnen; de anses lättare att utvärdera; och det faktum att man tidigt etablerade trenddata för dessa ämnen gör det naturligt att fortsätta mätningarna med relativt hög frekvens. En annan bidragande anledning är att de utvärderingsmetoder som utvecklats inom NAEP inte så lätt låter sig användas inom alla ämnen och ämnesdelar.

Enligt direktiven skall nationella kunskapsbedömningar genomföras årligen, vilket innebär att de årliga undersökningarna bör omfatta en grupp av ämnen. Viktiga frågor är då dels vilka ämnen som bör förekomma tillsammans, dels om eleverna skall delta i endast ett ämne, eller i flera.

Ur praktisk synpunkt framstår det som angeläget att systemet har en flexibilitet vad gäller möjligheter till val av vilka ämnen som skall utvärderas ett visst år. Det är inte självklart att periodiciteten skall vara densamma för alla ämnen och inte minst under uppbyggnadsperioden kan förutsättningarna att bygga upp och genomföra de nationella kunskapsbedömningarna variera över ämnena. Enligt detta resonemang bör ämnen i största möjliga utsträckning kunna kombineras fritt vid olika utvärderingstillfällen.

Samtidigt förhåller det sig så att det finns grupper av närliggande ämnen som med fördel skulle kunna behandlas tillsammans. Detta gäller inte minst blockämnena inom det naturorienterande området (fysik, kemi, biologi och teknik) och inom det samhällsorienterande området (historia, samhällskunskap, religion och geografi). Det är också möjligt att hävda att SO-området vid sidan av dessa fyra ämnen består av ett femte ämne, nämligen det samordnade SO-ämnet med en egen kursplan och ett eget betyg, liksom det även finns ett samordnat NO-ämne med en egen kursplan och ett eget betyg. Om inte ämnena inom de två blocken behandlas samtidigt försvåras möjligheterna avsevärt att utvärdera de två blockämnena. Det framstår därför som naturligt att SO-ämnena behandlas vid ett och samma tillfälle, liksom att NO-ämnena behandlas vid ett och samma tillfälle. Inom NU-03 behandlades de fyra SO-ämnena samtidigt och utvärderingen fokuserade även det övergripande SO-ämnet. Tre av NO-ämnena (fysik, kemi, och biologi, men inte teknik) redovisas i samma rapport, men utan särskilt fokus på det övergripande NO-ämnet.

Om eleverna deltar i kunskapsbedömningen av mer än ett ämne är det möjligt att länka ihop resultaten för de olika ämnena, och presentera dem på samma skala. Ytterligare ett skäl för varför det kan vara en vinst att låta elever delta i mer än ett ämne är att detta förbättrar möjligheterna till fördjupade analyser, genom att samvariationen mellan resultat inom olika områden kan studeras. Inom PISA får eleverna vid varje undersökningstillfälle besvara uppgifter inom läsning, matematik och naturvetenskap, vilket gjort det möjligt att bland annat belysa den stora betydelse som läskompetens har för prestationer inom matematik och naturvetenskap. Redan Härnqvist (1975) förslog att analyser i vilka läsresultat används som kontrollvariabel skulle vara av stort värde i analytiska sammanhang.

Det finns ytterligare skäl för varför det kan vara mindre ändamålsenligt att avgränsa elevernas deltagande till att omfatta endast ett ämne, och det är att mycket av skolans verksamhet fokuserar på teman som spänner över flera ämnen. Det kan också noteras att den nya lärarutbildningen fäster stort avseende vid tvärvetenskap och samverkan över de traditionellt definierade ämnesgränserna. Möjligheterna att inom kunskapsbedömningen inkludera uppgifter som är ämnesövergripande är givetvis sämre om fokus ligger på ett ämne i taget.

Eftersom definitionen av vad som är ett skolämne är starkt kopplat till läroplanerna, finns det också en risk att ändringar i läroplanerna medför behov av att ändra i utformningen av den nationella kunskapsbedömningen om denna utgår från de ämnen som finns vid en viss tidpunkt. Detta strider mot principen att ramverken för den nationella kunskaps-bedömningen skall vara stabila under längre tid. Även detta talar för att den nationella kunskapsbedömningen inte bör fokuseras på ett enskilt ämne i taget.

Även om det finns praktiska fördelar med att fokusera på ett ämne i taget pekar diskussionen ovan på att det ur flera synpunkter som avser innehåll, analyser, och trendmätning medför fördelar att snarare strukturera kunskapsbedömningen kring kluster av ämnen. Sådana kluster av ämnen kan skapas med utgångspunkt från flera olika principer. Ur analyssynpunkt skulle det vara en fördel att kombinera de ämnen som inkluderar mer generella färdigheter som läsning, skivning, och matematik med i stort sett alla andra ämnen, vilket dock inte låter sig göras om inte alla ämnesområden behandlas vid ett och samma tillfälle. En annan princip är att föra närliggande och i skolan samverkande områden samman, vilket framförallt har fördelar ur innehållsliga synpunkter.

Vid sidan av att vara innehållsligt närliggande, medför, som redan påpekats, förekomsten av ett övergripande SO-ämne och ett övergripande NO-ämne i sig att det är naturligt att föra samman de fyra SO-ämnena till en grupp, och de fyra NO-ämnena till en annan grupp.

Matematikämnet förs ofta också till NO-gruppen, och den centrala betydelse matematik har för åtminstone delar av det naturvetenskapliga området kan göra det viktigt att koppla dessa ämnen till varandra.

Det finns också skäl för att koppla samman de främmande språken (dvs de moderna språken tyska, franska och spanska, tillsammans med engelska). Anledningen till detta är att endast engelskämnet är obligatoriskt, medan de moderna språken är valbara och kan ersättas med andra språk. Detta innebär att resultaten delvis påverkas av selektions-effekter, och för att det skall vara möjligt att beskriva trender i kunskaps- och färdighets-utvecklingen måste det finnas tillgång till någon form av kontrollvariabel. Det är mest naturligt att använda resultaten i engelska som en sådan kontrollvariabel.

Ytterligare en grupp av ämnen utgörs av de praktisk-estetiska ämnena: Bild, Hem- och konsumentkunskap, Idrott och hälsa, Musik och Slöjd. Även om det kanske inte finns någon anledning att försöka skatta ett gemensamt resultat för hela denna grupp av ämnen, finns det intressanta möjligheter att formulera uppgifter som har beröring med mer än ett av dessa ämnen.

En annan grupp av ämnen är vad som skulle kunna kallas ”modersmålsämnen”: Svenska, Svenska2, Modersmål (tidigare Hemspråk), och Teckenspråk. Dessa ämnen skulle kunna behandlas som fristående ämnen, utan någon länkning mellan uppgifter inom de olika ämnena. En annan möjlighet är att de förs samman med någon annan grupp, som SO-ämnena eller de praktisk-estetiska ämnena.

Grupperingen av ämnen är givetvis också i hög grad avhängig antalet grupper, vilket i sin tur har att göra med periodiciteten för upprepning av kunskapsbedömningen liksom med urvalet av årskurser. Dessa frågor diskuteras nedan.

Årskurser och periodicitet

Det är vanligt att de nationella utvärderingssystemen omfattar åk 4 och åk 8. Ett argument för detta val är att dessa årskurser i många länder representerar slutåren i de två första stadierna inom utbildningssystemet. Detta gäller exempelvis NEMP. Där kombineras dessa argument med argument för en 4-årig periodicitet, vilket innebär att samma grupp av ämnen utvärderas i åk 4 och åk 8 för samma ålderskohort. Även TIMSS genomför utvärderingar i åk 4 och åk 8 vart fjärde år, och argumentet för denna periodicitet är detsamma som i NEMP, dvs att samma ålderskohort studeras i åk 4 och åk 8.

För Sveriges del skulle ett alternativ i stället kunna vara åk 5 och åk 9, eftersom dessa årskurser utgör slutår för ett slags ”stadier” inom utbildningssystemet, vilket bland annat illustreras av att mål finns formulerade för dessa årskurser, liksom att det finns nationella prov i dessa årskurser. Samtidigt utgör förekomsten av mål ett problem, eftersom ett nationellt kunskapsbedömningssystem inte skall styras av mer specifika läroplans-formuleringar och ändringar i dessa. Det finns givetvis också en risk för kollision mellan de nationella proven som alla elever genomför, och ett stickprovsbaserat nationellt utvärderingssystem om dessa ligger i samma årskurs. Inte minst skulle belastningen på vissa elever riskera att bli stor om den nationella kunskapsbedömningen skulle förläggas till vårterminen i åk 9. Dessa argument talar för att det är lämpligt att undvika vt åk 9, och istället välja årskurserna 4 och 8.

Det är dock inte självklart eller nödvändigt att samma årskurser skall väljas för alla ämnen, och det kan finnas skäl att variera årskurs/termin över olika grupper av ämnen. För de ämnen som eleverna börjar läsa sent, som de moderna språken, kan det t ex hävdas att det är mindre meningsfullt att undersöka dessa före åk 9, och en kompromiss skulle då kunna vara att lägga dessa på höstterminen i åk 9. Behov av samordning med inter-nationella studier kan också ge anledning att variera årskurserna för olika ämnen.

Inom NAEP varierar det upprepningsintervall enligt vilka olika ämnesområden återkommer i utvärderingarna för olika ämnen, med mycket frekventa utvärderingar av läsning och matematik (numera vartannat år), och mycket sporadiska utvärderingar av andra ämnen (t ex bild). Inom NEMP däremot är periodiciteten densamma för alla ämnen, nämligen fyra år. Som grundprincip bör gälla att den svenska nationella kunskapsbedömningen av alla grupper av ämnen upprepas med samma intervall, men att det skall gå att rucka på denna princip om det finns skäl för detta.

Ett skäl som brukar anföras för att inte upprepa nationella kunskapsbedömningar med alltför korta mellanrum, eller att mer än sporadiskt delta i internationella undersökningar, är att förändringar på populationsnivå i allmänhet är mycket små och svåra att upptäcka

(en mer omfattande diskussion om detta förs nedan, se sid 101). Även om detta ofta är fallet visar utvecklingen i Sverige från 1995 till 2003, med en nedgång i matematik-prestationerna med en effektstorlek om drygt 0,4 för åk 8 enligt TIMSS, på att denna regel inte är utan undantag. Det kan också invändas att om förändringarna är små men ändå betydelsefulla talar detta snarare för att kunskapsbedömningarna bör upprepas ofta snarare än sällan, eftersom detta förbättrar möjligheten att fånga upp den trendmässiga förändringen (se Grissmer et al., 2003).

Som redan påpekats varierar periodiciteten kraftigt inom NAEP, från två år och uppåt, vilket medför att möjligheterna att upptäcka förändringar i nivåer av kunskaper och färdigheter varierar kraftigt över olika områden. Som också redan nämnts är periodiciteten i NEMP och TIMSS fyra år. För PIRLS är upprepningsintervallet fem år. PISA upprepas vart tredje år, men ämnena återkommer som huvudområde endast vart nionde år.

Kostnaderna för den nationella kunskapsbedömningen påverkas givetvis i hög grad av upprepningsintervallets längd, och kostnadsaspekten talar mot att ha kort periodicitet. Att ha längre tidsintervall än fem år framstår som riskabelt med tanke på att detta begränsar möjligheterna att upptäcka förändringar i nivån av kunskaper och färdigheter. Som en kompromiss skulle en periodicitet på fem år kunna vara lämplig. Detta skulle också göra det naturligt att arbeta med fem grupper av ämnen, även om det givetvis är tänkbart att ha, exempelvis, fyra grupper av ämnen och att ha ett ”pausår” vart femte år.

Även om grundprincipen bör vara att samma intervall skall gälla för alla grupper av ämnen, är det inte heller alldeles självklart att detta måste gälla exakt på året. Om en mer direkt koppling skall göras mellan den svenska nationella kunskapsbedömningen och de internationella studierna blir det nödvändigt att variera periodiciteten över ämnen, eftersom denna varierar över studierna. Frågan om koppling till de internationella studierna diskuteras nedan.

Utnyttjande av internationella studier

Enligt uppdraget från regeringen gäller att:

> Den nationella utvärderingen bör så långt som möjligt utnyttja såväl de erfarenheter som följer av Sveriges deltagande i internationella uppföljningar som resultaten från dessa studier.

Även direktiven från Skolverket betonar att möjligheterna att utnyttja de internationella studierna skall utredas:

> Möjligheterna att utnyttja jämförande internationella kunskapsstudierna PISA, PIRLS och TIMSS som delkomponenter i ett framtida rullande utvärderingssystem skall diskuteras. Förslaget till ett rullande utvärderingssystem får dock ej utgå från att dessa delkomponenter finns att tillgå.

Som redan nämnts genomförs PISA vart tredje år inom områdena läsning, matematik och naturvetenskap, varvid ett av områdena i taget utgör huvudområde, och de övriga två utgör biområden enligt ett rullande system. I PISA används en åldersbaserad urvalsmodell och populationen utgörs av 15-åringar. I Sverige går huvuddelen av dessa i åk 9. Inom PISA fokuseras de kunskaper och färdigheter som antas vara betydelsefulla inom fortsatta studier och yrkesverksamhet.

Det framstår dock inte som möjligt att direkt använda PISA som en komponent i ett svenskt nationellt kunskapsbedömningssystem, och då framförallt av det skälet att den gruppering av ämnen som diskuterats ovan inte är förenlig med den gruppering som används inom PISA. Inte heller är periodiciteten i PISA förenlig med den som diskuterats ovan. Även om det sålunda framstår som omöjligt att låta resultaten från PISA ingå som en komponent i den nationella kunskapsbedömningen är givetvis resultaten från PISA av stort intresse som ett kompletterande underlag vid bedömningar av utvecklingen inom de tre områdena. Vid utvecklingen av ramverk för den nationella kunskapsbedömningen är det också av stort intresse att som en av utgångspunkterna använda de ramverk som utvecklats inom PISA.

PIRLS är en studie av läsfärdigheten hos yngre elever med fokus på åk 4. Den första undersökningen genomfördes år 2001, den andra kommer att genomföras år 2006, och PIRLS kommer sedan att upprepas vart femte år. Sverige deltog tillsammans med 34 andra länder i PIRLS 2001, och kommer även att delta i PIRLS 2006 tillsammans med ett 50-tal länder. Enligt den uppläggning som skisserats ovan skulle det inte föreligga några hinder att regelmässigt låta PIRLS utgöra en komponent i ett svenskt nationellt kunskapsbedömningssystem som indikator på läskompetens hos yngre elever. Detta skulle också medföra flera fördelar:

- En svensk trendlinje för åk 4 startades redan år 2001.
- Möjligheterna till jämförelse med ett stort antal andra länder ger ytterligare information som är av värde vid tolkningen av de svenska resultaten.
- Kostnaderna för utveckling av instrument, datahantering, analys och delar av rapporteringen delas mellan ett stort antal länder.
- Kostnaderna och belastningen på skolan blir mindre om den nationella kunskapsbedömningen kan utnyttja resultat från PIRLS, än om två undersökningar utförs mer eller mindre parallellt.

Deltagande i PIRLS är givetvis inte en förutsättning för bedömning av yngre elevers läskompetens i Sverige. Skulle det av någon anledning inte vara möjligt att delta i detta internationella projekt är det möjligt att med användning av i princip samma metoder som inom PIRLS utveckla egna instrument, och genomföra en egen undersökning.

Som nämnts ovan genomförs inom TIMSS vart fjärde år (1995, 1999, 2003, 2007, 2011, …) undersökningar av matematik och naturvetenskap i åk 4 och åk 8. Sverige deltog med åk 8 år 1995 och 2003, och avser att delta med åk 4 år 2007.

En möjlighet är att regelmässigt använda TIMSS i åk 4 och åk 8 vart fjärde år för att därigenom delvis täcka av områdena matematik och naturvetenskap inom den nationella kunskapsbedömningen. Ett problem är dock att de ramverk för matematik och naturvetenskap som TIMSS bygger på inte fullt ut kan förväntas sammanfalla med de ramverk som skulle konstrueras om en motsvarande svensk undersökning skulle byggas från grunden. Ett sätt att komma tillrätta med detta problem skulle kunna vara att i ett första steg skapa ett svenskt ramverk för matematik och naturvetenskap, och därefter göra en detaljerad jämförelse mellan ramverk och uppgifter för TIMMS å ena sidan och det svenska ramverket å den andra. En sådan jämförelse kommer att visa på tre utfall:

1. Områden som i samma utsträckning är representerade i TIMSS och i det svenska ramverket.
2. Områden som är orepresenterade eller underrepresenterade i TIMSS jämfört med det svenska ramverket.
3. Områden som är överrepresenterade i TIMSS jämfört med det svenska ramverket.

I bästa fall utgör utfall 1 resultatet av den största andelen av alla jämförelser, men det är givetvis rimligt att förvänta sig att även diskrepanser kommer att upptäckas. Vad gäller de områden som är underrepresenterade i TIMSS (utfall 2) är en tänkbar lösning att skapa kompletterande uppgifter, så att dessa tillsammans med TIMSS-uppgifterna svarar mot det svenska ramverket. De områden som är överrepresenterade i TIMSS (utfall 3) är svårare att hantera, därför att det av praktiska skäl inte är möjligt att ta bort uppgifter vid administrationen av undersökningen. Förmodligen ställer det sig också tekniskt komplicerat att vid skattning av förändring avstå från att inkludera uppgifter, även om denna möjlighet inte skall uteslutas. Om det inte skulle vara möjligt att tona ned de delar av TIMSS som är överpresenterade jämfört med det svenska ramverket är det dock givetvis möjligt att vid tolkningen av resultaten ta hänsyn till detta. Samma resonemang gäller givetvis även för andra internationella studier än TIMSS.

Utan en detaljerad analys av det slag som skisserats ovan är det i detta sammanhang svårt att dra bestämda slutsatser om möjligheten att utnyttja TIMSS som en komponent i det nationella kunskapsbedömningssystemet. Alternativet är givetvis att skapa ett nytt svenskt system från grunden, vilket i så fall skulle få en teknisk utformning som i mycket skulle likna TIMSS. Detta skall då kostnads- och funktionsmässigt jämföras med kostnad och funktion för ett TIMSS som är förstärkt med uppgifter inom de områden som är underrepresenterade i TIMSS. Om detta är möjligt kan i princip samma fördelar uppnås som nämndes i anslutning till diskussionen om PIRLS, förutom att en svensk trendlinje för åk 8 startades redan år 1995, och en trendlinje för åk 4 kommer att starta år 2007.

Om ett sådant ”förstärkt TIMSS” skall utgöra en av delarna i den nationella kunskapsbedömningen är det nödvändigt att för denna del använda en periodicitet om fyra år. Det är dock knappast ett problem att för detta område upprepa undersökningarna med periodiciteten fyra år, och använda femårig periodicitet inom andra områden.

Förslag till grundläggande design

Mot bakgrund av diskussionen ovan framstår följande gruppering av ämnen som ett möjligt förslag:

Praktisk-estetiska	**Främmande språk**	**Modersmål**	**Matematik/ NO**	**SO**
Bild	Engelska	Svenska	Matematik	Samhällskunskap
Hem- och konsumentkunskap	Franska	Svenska 2	Biologi	Historia
Idrott och hälsa	Spanska	Modersmål	Fysik	Religion
Musik	Tyska	Teckenspråk	Kemi	Geografi
Slöjd			Teknik	Integrerat SO
			Integrerat NO	

Det måste understrykas att denna gruppering är tentativ, och placeringen av enskilda ämnen kan givetvis diskuteras. Modersmål skulle exempelvis kunna föras till gruppen ”Främmande språk”. Det kan också noteras att det är stort och kostsamt arbete att utveckla ramverk och uppgifter för alla ämnen, vilket kan göra att implementeringen försvåras. Ett sätt att mildra detta problem är att inte ta med samtliga ämnen i första omgången, utan att lägga till ämnen successivt. Detta är dock bara möjligt för de grupper av ämnen där eleverna endast deltar i ett eller några ämnen (dvs de ämnen som förts samman under grupperna Praktisk-estetiska ämnen och Modersmålsämnen). En annan möjlighet att minska implementeringsproblemen är att första omgången endast fokusera på tidiga eller senare år. Ytterligare ett sätt att minska belastningen är att använda fler grupper av ämnen och en längre periodicitet, och eventuellt en varierad periodicitet för olika grupper av ämnen.

I tabellen nedan lämnas förslag till startår för kunskapsbedömningarna av de olika grupperna av ämnen. Bortsett från den eventuella kopplingen till de internationella undersökningarna är dessa godtyckligt valda. Under antagande att kunskapsbedömningen av matematik/NO genomförs i anslutning till TIMSS måste startåret bli 2011, med en fyrårig periodicitet. Detta innebär att två grupper av ämnen kommer att undersökas år 2011, medan ingen undersökning kommer att genomföras år 2012. För Modersmål föreslås startåret 2011, eftersom detta gör det möjligt att utnyttja PIRLS i åk 4. Det kan eventuellt även finnas anledning att koppla kunskapsbedömningen i SO till den internationella undersökningen International Civics and Citizenship Study (ICCES) för vilken datainsamlingen planeras till 2008.

	Främmande språk	**Praktisk-estetiska**	**SO**	**Modersmål**	**Matematik/ NO**
Periodicitet	5 år	5 år	5 år	5 år	4 år
Omgång 1	2008	2009	2010	2011	2011
Omgång 2	2013	2014	2015	2016	2015
Omgång 3	2018	2019	2020	2021	2019
Omgång 4	2023	2024	2025	2026	2023
Omgång 5	2028	2029	2030	2031	2028

Utformning av kunskapsbedömningarna

Nedan diskuteras mer i detalj hur bedömningarna av kunskaper och färdigheter kan utformas.

Uppgifts- och skalorienterade ansatser

Genomgången av de internationella erfarenheterna i Del I visar på en stor variation i de sätt på vilket kunskapsbedömningarna konstruerats, bedömts, analyserats och rapporterats. Även om det innebär ett visst mått av förenkling kan en distinktion göras mellan två olika ansatser till nationell kunskapsbedömning: en som kan betecknas som uppgiftsorienterad; och en som kan betecknas som skalorienterad. I den uppgiftsorienterade ansatsen ligger fokus i undersökningsdesign, uppgiftskonstruktion, analys, rapportering och den praktiska användningen av resultaten på en innehållsnära nivå, och den naturliga enheten är den enskilda uppgiften eller en grupp av likartade uppgifter. I den skalorienterade ansatsen ligger fokus på utveckling, analys och tolkning av skalor som representerar nivåer av prestationer inom olika innehållsliga domäner.

Givetvis skapas skalorna alltid från en större eller mindre uppsättning uppgifter, så i viss mening kan den skalorienterade ansatsen ses som en vidareutveckling av den uppgiftsorienterade ansatsen, genom att flera uppgifter aggregeras till en skala. Detta är dock ett alltför enkelt sätt att betrakta distinktionen mellan de två ansatserna. Ett skäl för detta är att de tenderar att vara associerade med olika tankemodeller och målsättningar. Ett annat skäl är att de olika empiriska angreppssätt som de två ansatserna implicerar eller tillåter får återverkningar på vilka typer av innehåll, uppgifter och resultat som de lyfter fram, och för vilka intressenter som resultaten är meningsfulla. Detta gör det rimligt att betrakta de två ansatserna som komplementära, snarare än att se den skalorienterade ansatsen som en påbyggnad på den uppgiftsorienterade ansatsen.

Vissa kunskapsbedömningssystem är lätta att klassificera i endera av dessa kategorier. NAEP var i den form som det ursprungligen introducerades av Tyler år 1969 ett uppgiftsorienterat system, men i och med att den moderna mättekniken togs i bruk vid 1980-talets början förvandlades NAEP till ett skalorienterat system. NEMP, som i stor utsträckning är baserat på Tylers idéer, är ett uppgiftsorienterat system, medan det franska BILAN-systemet är skalorienterat. De svenska nationella utvärderingarna (t ex NU-03) är uppgiftsorienterade. Det holländska PPON-systemet är dock inte lika lätt att klassificera. Visserligen används här IRT-teknik för att generera resultat på skalnivå. På grund av att mycket strikta krav på endimensionalitet upprätthålls inom PPON är dessa skalor dock mycket snävt definierade och i många fall omfattar de endast en uppgiftstyp. PPON kan sålunda betraktas som en blandning av ett uppgifts- och ett skalorienterat system. De internationella studierna (PISA, TIMSS, och PIRLS) är alla skalorienterade.

De två ansatserna har såväl sina för- som nackdelar. I anslutning till en genomgång av de viktigaste syftena och utmaningarna med ett nationellt kunskapsbedömningssystem

diskuteras dessa dels med avseende på validitet, dels med avseende på bestämning av förändringar i nivåer av kunskaper och färdigheter.

Validitet

Som framgår av diskussionen i kapitel 4 är frågor kring validiteten i beskrivningen av kunskaper och färdigheter (se sid 46) av fundamental vikt att beakta vid utformningen av ett nationellt kunskapsbedömningssystem. En huvudslutsats som drogs var att en förutsättning är att konsensus kan etableras kring ett ramverk som beskriver kunskaps- och färdighetsdomänerna i hela sin bredd och komplexitet. All erfarenhet visar också att ramverken implicerar en så stor mängd uppgifter att det är otänkbart att alla dessa genomförs av alla medverkande elever. Någon form av matrissamplingsteknik är därför nödvändig.

Matrissampling kan genomföras på flera olika sätt, och det föreligger stora skillnader mellan den uppgiftsorienterade och den skalorienterade ansatsen i detta avseende. Den viktigaste skillnaden är att i den skalorienterade ansatsen är det nödvändigt att se till att grupper av uppgifter som besvaras av olika elever överlappar för att det skall vara möjligt att överföra resultaten till den gemensamma skalan, medan det i den uppgiftsorienterade ansatsen inte finns något sådant behov. Tvärtom kan det visas att de säkraste skattningarna av det nationella medelvärdet i den uppgiftsorienterade modellen erhålls om varje elev löser var sin uppgift. Det är dock opraktiskt att driva matrissamplingen till denna extrema form. Inom NEMP används exempelvis den tekniken att tre varandra helt uteslutande grupper av uppgifter ges till tre parallella urval av elever. Mellan de tre uppgiftsgrupper som bjuds vid ett visst undersökningstillfälle finns sålunda inget överlapp, medan däremot ungefär en tredjedel av uppgifterna återkommer vid nästa undersökningstillfälle.

I den skalorienterade modellen krävs att det finns en direkt eller indirekt länk mellan alla uppgifter, vilket kan åstadkommas på en rad olika sätt, t ex genom s k ”balanced incomplete block design” (se sid 23). Detta innebär dock vissa restriktioner på uppgifternas utformning. Framförallt är det viktigt att de kan sammanföras till block som ställer samma tidskrav, och det är också viktigt att instruktioner och genomförandeformer är enkla och enhetliga. I den uppgiftsorienterade modellen kan däremot en uppgift utformas friare, utan hänsyn till andra uppgifters krav och egenskaper. Inte heller behöver instruktioner och genomförande här kringgärdas med samma restriktioner, vilket innebär att exempelvis gruppuppgifter är lättare att implementera i den uppgiftsorienterade modellen.

En av de mest diskuterade frågorna i samband med nationell kunskapsbedömning avser just uppgifternas validitet (se sid 53), varvid det ofta hävdats att inslaget av mer omfattande, produktiva, uppgifter (dvs uppgifter som man brukar beteckna som ”laborativa”, ”performance”, ”extended-constructive response (ECR)”, eller autentiska uppgifter) är alltför begränsat i de storskaliga kunskapsbedömningarna. Samtidigt har det också hävdats att denna typ av komplexa uppgifter inte lämpar sig för bruk i de

storskaliga undersökningssammanhangen, och framförallt inte då den skalorienterade ansatsen används.

Som framgår av redovisningarna i kapitel 4 av erfarenheterna från de stora ansträngningar som gjorts att inkludera sådana uppgifter i NAEP finns det visserligen positiva erfarenheter, men en stor mängd problem har också observerats, som låg svarsfrekvens, dålig reliabilitet, tidsbrist, stora kostnader för bedömningsarbetet, och klent informationsutbyte. En huvudslutsats är att denna typ av uppgifter inte är speciellt lämplig att använda i den skalorienterade ansatsen (Forsythe et al. 1996; Pellegrino et al., 1999). Inom NEMP däremot tycks erfarenheterna av de komplexa, produktiva, uppgifterna huvudsakligen vara positiva, vilket pekar på att dessa uppgifter är lättare att hantera inom den uppgiftsorienterade ansatsen. En anledning till detta är den relativt begränsade storleken på de elevurval som används inom NEMP.

Denna diskussion pekar på att det i den skalorienterade ansatsen ställer sig betydligt svårare att med gott utbyte använda mer komplexa och omfattande uppgifter som ger möjligheter för eleverna att själva eller i grupper producera mer omfattande svar. Den främsta anledningen till detta är att denna ansats lägger restriktioner på utformning och genomförande av uppgifter och hantering av elevsvar, vilket inte i samma utsträckning är fallet med den uppgiftsorienterade ansatsen.

Trendbestämning

Det viktigaste syftet med det nationella kunskapsbedömningssystemet är att följa eventuella förändringar i nivån av kunskaper och färdigheter över tid. Genomgången av erfarenheter från trendmätning inom NAEP och de övriga nationella utvärderingssystemen visar att flera olika sätt använts:

- Under den urspungliga NAEP-designen genomfördes studierna av trend på uppgiftsnivå eller för kluster av uppgifter (se sid 20), och så görs också inom NEMP. I båda dessa fall har man publicerat en viss andel av uppgifterna i anslutning till resultatredovisningen.
- Inom ramen för LTT NAEP har man använt samma uppgiftsuppsättningar sedan 1989, vilket gjort det möjligt att studera förändringar i resultatfördelningar för olika undergrupper. Dessa uppgifter har hållits hemliga vilket varit möjligt därför att LTT NAEP endast utgör en mindre del av hela NAEP-systemet.
- Som visats inom NAEP och de internationella studierna är det också möjligt att inom ramen för väldefinierade ramverk successivt byta ut uppgifter, och genom användning av flergrupps IRT modeller studera trender.

Även om erfarenheterna från trendmätning inom LTT NAEP genom återanvändning av samma instrument delvis är goda, har under senare tid förändringar gjorts som innebär att i princip samma tekniker anammats som inom huvud NAEP. En huvudanledning till detta är att det visat sig nödvändigt att successivt förnya uppgifterna. Även om det är enkelt och kostnadseffektivt att vid varje tillfälle använda samma uppgifter är detta inte hållbart i längden, och i synnerhet inte inom områden med en snabb kunskapsutveckling. En

annan nackdel med att använda samma uppgifter är att detta förutsätter att uppgifterna kan hållas hemliga. Även om det är möjligt att hemlighålla en del av uppgifterna från ett tillfälle till ett annat kräver ett svenskt nationellt kunskapsbedömningssystem en sådan öppenhet att det inte är rimligt att bygga trendstudier på att instrumenten i sin helhet skall hållas hemliga. Sålunda återstår som huvudansatser den uppgiftsorienterade modellen som används inom exempelvis NEMP, och den skalorienterade modell som används inom NAEP och de internationella studierna.

Efter att NAEP under de första 10 åren mött en hel del problem vid användning av IRT-baserade metoder för trendmätning framstår detta nu som en väl etablerad teknik. En viktig förutsättning för att korrekta resultat skall nås är dock att förändringar i uppgifter och procedurer genomförs med stor försiktighet, och att trendmätningarna genomförs på grundval av ett väl definierat ramverk som i huvudsak är stabilt över tiden. Enligt den policy för utveckling av ramverk som formulerats av NAGB skall ramverk och test-specifikationer vara hållbara under minst 10 år. Under denna tidsrymd skall det sålunda inte vara några problem, bortsett från sådana som är av teknisk och empirisk art, att genomföra trendmätningar. I vilken utsträckning det är möjligt att göra trend-bestämningar som omfattar även mer eller mindre reviderade ramverk är såväl en konceptuell som en empirisk fråga.

Som ett exempel kan nämnas ramverken för läsning inom NAEP. År 1992 fastställdes ett ramverk för utvärdering av läsning i åk 4, 8 och 12, vilket varit i huvudsak oförändrat sedan dess och på grundval av vilket en trendlinje bestämts för ett halvdussin mät-tillfällen. NAGB (2005) har emellertid preliminärt fastställt ett nytt ramverk för läsning, vilket skall gälla från år 2009. Det gamla och det nya ramverket specificerar i huvudsak samma texttyper (”literary texts” och ”informational texts”), och även de typer av processer som skall utvärderas sammanfaller till stor del. En avgörande skillnad är dock att det nya ramverket specificerar att även vokabulär skall prövas, genom att uppgifter som avser förståelsen av en del av de ord som ingår i texterna inkluderas. Framförallt på grund av detta fastslår NAGB (2005) att en ny trendlinje skall etableras med det nya ramverket och att den tidigare trendlinjen skall avslutas med 2007 års undersökning.

Samtidigt föreslås att en specialundersökning skall genomföras med syfte att undersöka i vilken utsträckning införandet av vokabuläruppgifter påverkar den totala poängen. Det är givetvis tänkbart att en sådan undersökning kan komma att visa att införandet av vokabuläruppgifter inte förhindrar att den gamla och den nya trendlinjen kan förenas. Eftersom LTT NAEP kommer att fortsätta att undersöka trenden vad gäller utvecklingen av läsfärdigheter är inte heller behovet lika starkt inom detta område som inom andra områden att upprätthålla stabilitet och kontinuitet i ramverken. Om så inte varit fallet hade en möjlighet varit att i detta fall söka bevara trendlinjen genom att skatta poäng på en skala som inte inkluderar vokabuläruppgifterna, och därtill bestämma en separat vokabulärskala.

Under förutsättning att ramverken kan hållas i huvudsak oförändrade finns sålunda goda möjligheter att med de skalorienterade metoderna undersöka förändringar över tid, och

även då förändringar sker av ramverken kan under vissa förutsättningar samma trendlinje skattas även fortsättningsvis.

Den uppgiftsorienterade ansatsen till trendmätning fokuserar på identiska uppgifter. Inom NEMP används samma grundläggande teknik för att studera förändring över tid som ursprungligen brukades inom NAEP, nämligen att administrera samma uppgifter som använts tidigare, och att sedan undersöka förändring på uppgiftsnivå. Som grundprincip återupprepas ca 30 % av de uppgifter som användes vid det föregående undersökningstillfället.

Detta är i princip en enkel teknik, men den vidlåds också av begränsningar. För det första kan den leda till svårigheter då syftet är att etablera trendlinjer över längre tid. För att detta skall kunna ske är förutsättningen att samma uppgifter upprepas vid varje tillfälle, vilket i sin tur förutsätter att uppgifterna hålls hemliga. Eftersom trycket är starkt på att redovisa åtminstone vissa av de använda uppgifterna är det dock knappast möjligt att hemlighålla alla de uppgifter som används för trendmätning. Inom NEMP har detta som konsekvens att trendbestämning ofta endast kan göras genom jämförelse med närmast föregående tillfälle. För läsförståelse kunde sålunda trendbestämning för åk 4 göras med tre texter mellan 2000 och 2004, medan det inte fanns någon text som kunde användas för jämförelse mellan 1996 och 2004. För högläsning var två texter gemensamma mellan 2000 och 2004, medan en text förekom både 1996 och 2004. Det faktum att trendlinjen i huvudsak baseras på jämförelser mellan successiva undersökningsomgångar gör att denna kan bli instabil, vilket problem ytterligare försvåras av att jämförelserna baseras på ett mycket begränsat antal uppgifter. De förändringar som observeras för en enstaka text kan sålunda vara i hög grad beroende av om det innehåll texten behandlar har uppmärksammats i undervisningen eller media.

Ett annat problem är att den uppgiftsorienterade ansatsen till trendmätning i första hand lämpar sig för redovisning på uppgiftsnivå i termer av procent eller medelvärden. Detta kan leda till en nyanserad och mångfasetterad beskrivning av utvecklingen inom ämnesområdet, vilken dock inte ger den översiktliga information som efterfrågas på den nationella nivån. Denna uppgiftsnära redovisning lämpar sig inte heller för beskrivning av skillnader i resultat och resultatfördelningar för olika undergrupper, och inte heller för förklaringsinriktade analyser. I den mån det är möjligt att använda samma uppsättning uppgifter över tid, över vilka resultaten kan aggregeras till exempelvis en summa, minskar dessa olägenheter.

Slutsats

Diskussionen ovan leder till slutsatsen att den uppgiftsorienterade ansatsen vidlåds av begränsningar när syftet är att göra trendmätningar. Detta har framförallt sin grund just i att fokus ligger på enskilda uppgifter, vilket innebär begränsningar i möjligheterna att göra jämförelser som går över mer än två tidpunkter, och att resultaten kan bli beroende av utfallet för enstaka uppgifter. Den skalorienterade ansatsen har inte dessa begränsningar, eftersom aggregering där sker över många uppgifter. En förutsättning för

att denna ansats skall fungera är dock att ramverk och testspecifikationer är i huvudsak konstanta.

Samtidigt måste också konstateras att den skalorienterade ansatsen implicerar betydligt strängare restriktioner på vilka typer av uppgifter och bedömningssituationer som kan användas. Medan mer komplexa, produktiva, uppgifter utan särskilda problem kan användas i den uppgiftsorienterade ansatsen, är det av tekniska, praktiska och ekonomiska skäl svårt att använda dem i den skalorienterade ansatsen. Av validitetsskäl är det dock nödvändigt att det finns utrymme även för mer komplexa uppgifter i ett svenskt nationellt kunskapsbedömningssystem.

Det är därför inte möjligt att fastslå att systemet skall utformas antingen som ett renodlat skalorienterat eller ett renodlat uppgiftsorienterat system, utan dessa ansatser bör kombineras på ett sådant sätt att fördelarna hos de olika ansatserna kan tas till vara. Fördelarna med den skalorienterade ansatsen när det gäller trendmätning och beskrivningar av fördelningar för olika undergrupper måste sålunda utnyttjas, och fördelarna med de mer komplexa uppgiftstyperna vad gäller mer ingående och mångfacetterade beskrivningar av elevernas kunskaper och färdigheter måste tas till vara.

Den modell som här skisseras har stora likheter med den multimetodansats som Pellegrino et al. (1999) föreslog som en vidareutveckling av NAEP. Den sammanfaller också i mycket med förslaget från Forsyth et al. (1996) att differentiera NAEP dels i ett ”core NAEP” med huvudsyfte att göra nivå- och trendbestämningar, dels i specialstudier med syfte att göra mer mångfasetterade kunskapsbedömningar inom olika område.

Distinktionen mellan de skal- och uppgiftsorienterade ansatserna är långt ifrån entydig, och för att minska risken för otydlighet i kommunikationen finns det anledning att något närmare precisera vad som skulle kunna vara en rimlig innebörd av dessa begrepp i samband med ett svenskt kunskapsbedömningssystem.

Eftersom NAEP har relativt högt ställda ambitioner att inkludera komplexa uppgifter utgör inte detta system ett bra exempel på det slags skalorienterat system som här är aktuellt. Konkreta och förebildliga exempel skulle snarare vara de internationella studierna PISA, PIRLS och TIMSS. Dessa har väl utvecklade ramverk med stöd för trendbestämning, och de definierar breda och omfattande urval av uppgifter, vilka praktiskt hanteras inom ramen för matrissamplingsdesigner. Uppgifterna är dels flervalsuppgifter, dels uppgifter som kräver ett kort svar (från något ord till några meningar), och där de senare kräver en bedömningsinsats, vilken understöds av detaljerade bedömningsanvisningar. Även om uppgifterna inte kräver omfattande produktion har de dock ofta hög ambitionsnivå när det gäller att fånga även djupare förståelsenivåer och mer komplex problemlösning.

Ett konkret exempel på en förebildlig uppgiftsorienterad ansats ges av NEMP, och i NU-03 finns också goda exempel på komplexa uppgifter med en omfattande elevproduktion. Det framstår givetvis som angeläget att dra nytta av de erfarenheter och resultat som vunnits inom det svenska nationella utvärderingsprojektet.

Förslaget är sålunda att dessa två ansatser skall kombineras till en integrerad helhet. Båda ansatserna kan förväntas förekomma inom samtliga ämnesområden, även om tyngdpunkten också kan förväntas vara olika. Den mer preciserade fördelningen av kunskapsbedömningsinsatser över de två ansatserna görs vid utvecklingen av ramverket, och formerna för detta diskuteras nedan.

Utformning av ramverk för kunskapsbedömningarna

Ramverket är den nationella kunskapsbedömningens primära styrdokument och i detta preciseras såväl de innehållsliga bestämningarna av vad som är viktigt inom området, som de grundläggande principerna för konstruktion av uppgifter och uppläggning av kunskapsbedömningen.

Alla levande kunskapsområden är fyllda av motsättningar kring vad som är viktigt, hur kunskap skall visas, och så vidare. Ett ramverk måste därför byggas upp i en process där olika ståndpunkter konfronteras och diskuteras, med målet att nå fram till konsensus. En sådan process är tidsödande och inom NAEP avsätts ofta flera år till utvecklingen av ett ramverk.

Ramverket för ett visst område skall definiera de kunskaper och färdigheter som skall bedömas; precisera fördelningen av uppgifterna över olika innehålls- och processdimensioner; och ange vilka typer av uppgifter som skall användas. Ramverket skall också ange fördelningen och arten av uppgifter som skall användas för konstruktion av skalor, respektive användas på uppgiftsnivå.

Utvecklingen av ramverket skall ske enligt följande:

- Utvecklingsprocessen skall genomföras i en brett sammansatt grupp bestående av lärare, elever, föräldrar, skolledare, och ämnes- och läroplansexperter, där olika synsätt finns representerade och där frågor kring innehåll och uppläggning av kunskapsbedömningen utsätts för en omfattande och mångsidig diskussion. Synpunkter på preliminära versioner av ramverket skall inhämtas från externa expert- och lekmannagrupper. Utvecklingsprocessen skall ske på ett öppet och balanserat sätt, och skall inte på ett otillbörligt sätt kunna påverkas av olika intressegrupper.
- Vid utvecklingen av ramverket skall hänsyn tas till aktuellt innehåll i undervisningen som det formuleras i nationella och lokala läroplaner, aktuell forskning om inlärning och undervisning, och den önskvärda utvecklingen av elevernas kunskaper och färdigheter inom området. I inledningsfasen av arbetet skall de grundläggande frågorna och problemen identifieras i en ”frågeställningsuppsats”, som sammanfattar aktuell forskning och diskussion. Under arbetet skall hänsyn tas till information i många olika källor såsom läroplaner, nationella och internationella ”standards” inom olika undervisningsområden, forskningsresultat, internationella jämförande undersökningar och andra kunskapsbedömningar.

- Resultatet av utvecklingsprocessen skall presenteras i ett dokument som beskriver ramverket, ett som anger uppgiftsspecifikationer, och ett tredje dokument som beskriver vilka bakgrunds- och kontextvariabler som skall samlas in för det aktuella ämnet. Dessa tre dokument skall vara färdigställda så att tillräcklig tid finns för att konstruera uppgifterna innan utprövningen av dessa skall påbörjas. Specifikationsdokumentet skall innehålla: a) detaljerade beskrivningar av innehålls- och processdimensionerna, med angivelse av den relativa vikten för olika innehålls- och processkategorier; b) uppgifts- och svarstyper; c) rättnings- och bedömningsprocedurer; d) administrationsprocedurer; och e) exempel på uppgifter, inklusive bedömningsanvisningar. Ramverk och specifikationer skall normalt vara giltiga i minst 10 år.

Statistiska aspekter

En viktig aspekt på utformningen av kunskapsbedömningssystemet avser dimensionering av urval av personer och uppgifter, bland annat eftersom detta är av avgörande betydelse såväl för systemets kostnad som för dess funktion. Alla stickprovsbaserade uppföljnings-system har den begränsningen att skattningar av, exempelvis, förändringar över tid, eller av skillnader mellan olika undergrupper i populationen, vidlåds av en statistisk osäkerhet. I värsta fall är den statistiska osäkerheten så stor att det inte är möjligt att upptäcka ens mycket betydelsefulla förändringar eller skillnader med en viss uppläggning av uppföljningssystemet. Det är därför nödvändigt att utforma systemet så att det har förmåga att fånga upp förändringar och skillnader av den storleksordning som kan bedömas vara betydelsefull. Olika aspekter av detta problem diskuteras nedan, med särskilt fokus på den skalorienterade ansatsen, eftersom det i huvudsak är den för vilken diskussionen över huvud taget äger relevans.

Den första frågan som måste diskuteras är hur stora skillnader som skall kunna upptäckas. Inom statistiken används begreppet ”effektstorlek” för att ange storleks-ordningen på en förändring eller skillnad mellan två mätvärden. Detta anges på en standardiserad skala där värdet 0 anger avsaknad av effekt, och värden över 0.8 anger en stark effekt. Enligt de tumregler som etablerats är effekter i intervallet 0,2-0,3 att betrakta som svaga, och ofta anses effekter under 0,2 som betydelselösa.

Bedömningen av vilken effektstorlek som är av intresse kan dock inte göras på ett schabloniserat sätt utan måste ta hänsyn till det fenomen som undersöks. Förändringar över tid på nationell nivå är i allmänhet relativt små, och enligt erfarenheterna från NAEP är det relativt sällsynt med förändringar som är större än 0,1 över successiva mättillfällen (Grissmer et al., 2003). Anledningen till detta är att det är många faktorer som påverkar resultat på den nationella nivån. Samtidigt innebär den stora mängden elever som resultaten omfattar att även relativt små förändringar får betydelsefulla implikationer. I synnerhet om förändringen är trendmässig är det därför angeläget att kunna upptäcka även förändringar som i andra sammanhang skulle betraktas som små.

Som ett exempel kan nämnas att IEA's "Ten Year Trend Study" av förändringar i läsfärdigheten i åk 3 mellan år 1991 och 2001 visade på en nedgång i läsfärdigheten i Sverige som hade en effektstorlek på 0,16 (Martin, Mullis, Gonzalez, & Kennedy, 2003). Om nedgången varit linjär under den studerade 10-årsperioden betyder detta att effekten per 5-årsperiod var 0,08. Den totala nedgången i läskompetens är av stor praktisk betydelse, och det hade varit angeläget att kunna upptäcka denna tidigt. Även detta exempel pekar sålunda på att det är nödvändigt att med god sannolikhet kunna upptäcka förändringar med effektstorlekar som uppgår till 0,10.

Stickprovsstorleken är en de faktorer som har störst betydelse när det gäller möjligheten att upptäcka en skillnad med en viss effektstorlek. Då den önskade effektstorleken är bestämd på förhand är det också lätt att beräkna vilken stickprovsstorlek som behövs för att med viss sannolikhet kunna upptäcka en skillnad. Om vi exempelvis antar att vi med 80 % sannolikhet på 5 % nivån vill kunna upptäcka en effekt som uppgår till 0,1 behövs en stickprovsstorlek om 1570 elever vid var och en av de två mätningarna.

Denna uppskattning av vilken minimal stickprovsstorlek som krävs bygger dock på en rad antaganden, vilka i allmänhet inte är uppfyllda, och i själva verket medför olika omständigheter att den faktiska stickprovsstorleken behöver vara flerdubbelt större:

- Beräkningarna bygger på antagandet att de individer som ingår i urvalet är valda på ett oberoende sätt. Av såväl praktiska som designmässiga skäl görs emellertid urval och mätning praktiskt taget alltid med hela skolor eller klasser (s k klusterurval). Eftersom de elever som finns inom en skola eller klass prestationsmässigt tenderar att likna varandra mer än ett slumpmässigt urval av elever leder detta till en informationsförlust. Det effektiva antalet elever i stickprovet är därför lägre än summan av antalet elever i de olika skolorna och klasser. Det går att kompensera för detta genom att öka antalet individer i klusterurvalet, och ofta krävs en flerdubbling av antalet individer för att nå samma precision som då urvalet görs på individnivå.
- Det finns ofta anledning att använda stratifierade urvalsdesigner, i vilka olika individer i populationen har olika sannolikhet att komma med i urvalet, men den enkla beräkningen ovan bygger på antagandet att alla individer har samma sannolikhet att bli valda. Ett skäl för stratifiering är att öka precisionen i skattningarna, vilket gör det möjligt att uppnå en viss bestämd styrka med en mindre stickprovstorlek. Ett annat skäl för stratifiering är att göra jämförelser möjliga för små undergrupper i populationen (t ex elever i friskola). Genom att låta elever i sådana undergrupper ingå med en högre andel än vad som motsvaras av deras andel i populationen kan meningsfulla jämförelser göras även för små men intressanta grupper. Vid sådan användning av stratifiering krävs ofta ett större totalt urval än då stratifiering inte används. Stickprovets storlek, liksom stratifieringsdesignen, måste givetvis utformas så att det är möjligt att kunna upptäcka effekter av en viss preciserad storlek både för det totala materialet, och för de olika undergrupper som är av intresse.
- Beräkningen utgår också från antagandet att mätinstrumenten är felfria, alternativt att det finns en viss konstant mängd mätfel. Detta är dock ett förenklat synsätt, och i synnerhet då eleverna besvarar en begränsad mängd uppgifter i en

matrissamplingsdesign kan mätfelet bli av betydande storleksordning. En ökad omfattning av mätfelet måste kompenseras genom ett större urval av elever, och alternativt medför mindre mätfel möjlighet att minska urvalet. Då urvalsdesignen diskuteras är det därför nödvändigt att ta hänsyn till både mätfel och urvalsfel.

- Beräkningarna utgår också från att det är möjligt att få data för alla de elever som ingår i urvalet. I praktiken är detta dock aldrig möjligt, på grund av partiellt eller totalt bortfall. Om bortfallet är slumpmässigt är det möjligt att kompensera för ett förväntat bortfall genom att öka urvalsstorleken i motsvarande mån. Bortfall är dock sällan slumpmässigt, och i synnerhet om bortfallsfrekvensen för olika undergrupper ändras över tid kan detta medföra stora problem vid studium av förändringar.

Dessa komplikationer gör det till en grannlaga uppgift att fastställa en optimal storlek på urvalet, i synnerhet som detta innebär ställningstagande såväl till för vilka undergrupper som meningsfulla jämförelser kommer att kunna göras, som till utformningen av mätinstrumenten. Det faktum att relativt stora urval krävs gör det angeläget att inventera olika möjligheter att öka precisionen i undersökningsdesignen. Nedan diskuteras olika sätt att utforma undersökningsdesignen så att möjligheterna att upptäcka förändringar och skillnader blir så stora som möjligt

Mätprecision

Som påpekats ovan medför mindre mätfel i urvalsinstrumenten att det är möjligt att minska urvalet av antal elever. Mätfelen bestäms till största delen av hur många uppgifter en viss elev besvarar, vilket innebär att det finns en utbytbarhet mellan den mängd tid varje elev lägger på att besvara uppgifter, och det antal elever som bör ingå i urvalet. Detta innebär att den totala belastningen på skolsystemet som den nationella utvärderingen innebär i princip kan bli densamma om ett större antal elever var och en besvarar ett mindre antal uppgifter, som om färre elever besvarar ett stort antal uppgifter. Eftersom varje deltagande elev också skall besvara ett antal bakgrundsfrågor, och det dessutom finns en del andra ”fasta tidskostnader” för varje medverkande elev (t ex i form av information om undersökningen), torde det dock vara effektivare att använda ett större antal uppgifter för varje elev, och ett mindre urval av elever.

Denna utbytbarhet mellan den tid varje elev lägger på undersökningen och antalet elever som ingår i undersökningen illustreras är särskilt tydlig vid användning av matris-sampling. Antag att beräkningar visar att det behövs 400 elevsvar för varje uppgift för att den statistiska säkerheten skall vara tillräcklig. Om det totalt finns 160 uppgifter med en sammanlagd testtid om 320 minuter kan vi uppnå detta på många olika sätt, t ex genom att vi låter 800 elever besvara 80 uppgifter var under 160 minuter, eller genom att vi låter 1600 elever besvara 40 uppgifter var under 80 minuter.

Inom NAEP har man valt att låta minimera omfattningen av varje elevs medverkan. Vid sidan av ett formulär med frågor om elevens bakgrund och om utformningen av undervisningen, besvarar i allmänhet varje elev ett häfte med frågor under ca 40 minuter. Det finns flera skäl för varför det amerikanska nationella utvärderingssystemet utformats

på detta sätt. För det första innebär detta en markering av att de individuella resultaten inte är av intresse, eftersom dessa är så osäkra att det inte går att göra några uttalanden om en viss elevs resultat på grundval av så begränsade observationer. För det andra innebär det frivilla deltagandet, som inte heller är av någon betydelse för vare sig eleven, klassen, läraren eller skolan, att motivationen till mer omfattande insatser är begränsad. Erfarenhetsmässigt leder också en belastning av uppgifter som kräver mer än 40 minuter till svarsbortfall. För det tredje används bakgrundsuppgifterna som s k kolateral-information vid skattning av resultaten med hjälp av den betingade MML-tekniken, vilket leder till en förbättrad precision.

I Sverige är dock behovet att markera att de deltagande elevernas resultat inte skall användas för värdering av eleven, klassen, läraren eller skolan mindre än i USA, på grund av att användningen av provresultat för olika former av ansvarsutkrävande här inte är lika utbredd. Möjligheterna att öka precisionen genom att utnyttja bakgrundsinformation som kolateralinformation är också mer begränsad i Sverige än i USA. I USA kan mer än 50 % av variationen i resultaten förklaras från kännedom om bakgrundsvariabler, medan motsvarande siffra i Sverige inte torde överstiga 20 %. Med tanke på de tekniska komplikationer som i analyssteget följer av användning av kolateralinformation finns det också skäl att noga överväga om sådan information skall användas i skattningsmodellen.

Mot bakgrund av dessa skillnader mellan de amerikanska och de svenska förhållandena framstår det som lämpligare att i ett svenskt nationellt utvärderingssystem öka precisionen genom att låta varje elev medverka i större omfattning, och att inte använda bakgrundsinformation vid skattning av elevresultaten. Att här precisera vilken omfattning av medverkan som är optimal för varje elev låter sig inte göras. Det kan dock noteras att i NU-03 medverkade många elever i upp till 10 lektionstimmar, och en så omfattande medverkan torde inte behöva bli aktuell. Det torde vara rimligare att räkna med ca 3-5 lektionstimmar per elev.

Användning av registerbaserad information

Som påpekades ovan har bakgrundsinformation en central roll vid resultatskattningarna inom NAEP, vilket också gäller de internationella studierna. Inom PISA används dock bakgrundsinformationen endast vid skattning av individernas plausibla värden, och inte vid bestämning av parameterestimaten på nationell nivå. Eftersom svaren på enskilda frågor ofta vidlåds av en hög grad av osäkerhet och det finns en stor mängd frågor genomförs i allmänhet en reducering av den stora mängden variabler till ett mer begränsat antal principalkomponenter, vilka sedan används i beräkningsarbetet.

Denna teknik ökar precisionen i skattningarna i den mån det finns ett samband mellan bakgrundsinformationen och elevresultat och i allmänhet är detta samband betydande. Även för Sveriges del finns ett sådant samband, även om det är lägre än vad som är fallet i många andra länder, vilket framförallt har sin grund i att variationen i resultat mellan skolor är lägre i Sverige än i många andra länder.

Användning av kolateralinformation som samlats in direkt från eleverna är dock förknippad med flera problem. Som redan nämnts kan användning av en viss uppsättning variabler både vid MML-skattningen och vid analysen av olika förklaringsfaktorers betydelse leda till tvivel- eller felaktiga resultat i analyssteget. Ett annat problem är att svaren på enskilda frågor ofta är av mycket låg kvalitet, vilket exempelvis är fallet med uppgifter om föräldrarnas utbildning och yrke, och i synnerhet då dessa inhämtas från yngre elever. Detta utgör ett stort problem, särskilt då dessa variabler används i analysarbete som kontroll- eller förklaringsvariabler. Ytterligare ett problem med de uppgifter som inhämtas från eleverna är att dessa inte föreligger för de elever som av ett eller annat skäl inte deltar i undersökningen, och som sålunda utgör bortfall.

Till skillnad från många andra länder finns dock i Sverige under vissa omständigheter möjlighet att utnyttja register som källa till information om bakgrundsvariabler. Uppgifter om föräldrarnas utbildning och nationella bakgrund finns sålunda tillgängliga för de flesta elever, och kvaliteten på denna information är betydligt bättre än kvaliteten på de uppgifter som eleverna själva lämnar. Den finns också tillgänglig för de elever som inte besvarar något frågeformulär.

Under förutsättning att personnummer finns tillgängliga för de elever som valts ut att ingå i undersökningen kan SCB skapa en s k ”nyckeldatabas”, vilket innebär att information från olika register (t ex SUN-registret och RTB) läggs samman med information som samlas in med hjälp av frågeformulär och prov. För analys levereras sedan ett avidentifierat material.

Den registerbaserade informationen kan i princip användas som kolateralinformation i MML-skattningarna men som redan påpekats finns det skäl som talar mot detta. Det finns dock andra, och bättre, sätt att utnyttja denna information för att förbättra precisionen i skattningarna. För det första kan den användas för att undersöka bortfallets omfattning och art. Eftersom bortfall utgör en av de viktigaste felkällorna av såväl systematisk som slumpmässig art vid surveyundersökningar är detta viktig information. För det andra är det också möjligt att på statistisk väg korrigera för de eventuellt snedvridande effekter som bortfallet kan ha, varvid exempelvis programmet Mplus (Muthén & Muthén, 2004) kan användas. Detta görs lämpligen efter att de individuella plausibla värdena skattats.

Urvalsmodeller

Det är rimligt att i detta sammanhang använda ett flerstegsurval av det slag som används i NAEP och de internationella undersökningarna. I första steget väljs då skolor, eventuellt inom strata, enligt PPS-principen (”probability proportional to size”). Därefter görs urval av en, några eller alla klasser inom skolan. Urval av elever inom klasser kan också vara aktuellt, men i de fall en matrissamplingsdesign används är det lämpligt att låta alla elever i klassen ingå, och slumpmässigt distribuera de olika häftena till klassens elever.

Som påpekats ovan kan olika former av stratifierade urvalsmodeller användas för att öka precisionen i skattningarna, givet en viss urvalsstorlek. Det finns anledning att låta

exempelvis SCB utreda närmare vilka stratifieringsvariabler som skulle vara mest användbara för detta ändamål (t ex tätortsgrad, kommuntyp, skolstorlek).

En annan, kompletterande, möjlighet är att den andra och de efterföljande mätningarna genomförs vid samma skolor som ingick i den ursprungliga mätningen. Fördelen med en sådan ”fix-skol” design är att variationen mellan skolor inte påverkar bestämningen av förändring över tid, vilket är gynnsamt ur precisionssynpunkt. Nackdelen med en sådan design är dock att vetskapen om att skolan kommer att få återbesök inom samma ämnen som undersöktes tidigare kan medföra en ökad fokusering på just dessa ämnen, med ty åtföljande svårighet att bestämma i vilken utsträckning detta påverkat resultaten. Givet att huvudsyftet med den nationella kunskapsbedömningen är att studera förändring över tid är detta en så avgörande nackdel att denna modell inte bör användas.

Slutsats

På nationell nivå är även förändringar som i andra undersökningssammanhang betraktas som små betydelsefulla. Ett nationellt utvärderingssystem bör med god säkerhet kunna upptäcka förändringar över tid som har effektstorlek 0,1. För att detta skall vara möjligt måste urvals- och mätfelen hållas på låg nivå, liksom även bortfallet. Små urvalsfel kan komma att kräva stora urval av elever, och små mätfel kräver att varje elev genomför relativt många uppgifter. Eftersom både antalet elever och antalet uppgifter påverkar kostnaderna för undersökningen riskerar dessa att bli höga, om inte ansträngningar görs att optimera undersökningsdesignen. Sådana möjligheter finns i att utnyttja registerdata kring elevernas bakgrund för att kontrollera för bortfall, att optimera urvalsmodellen, och att balansera antal elever och antal uppgifter per elev.

Exempel på utformning av kunskapsbedömningar

I syfte att konkretisera diskussionen diskuteras nedan exempel på möjliga utformningar av kunskapsbedömningar. Särskilt diskuteras den samordnade användningen av den skal- och den uppgiftsorienterade ansatsen, och det eventuella utnyttjande av de internationella studierna.

En utgångspunkt tas i matematik och de naturorienterande ämnena. Det har ovan föreslagits att dessa ämnen bör behandlas vid samma utvärderingstillfälle, bland annat av det skälet att en samlad kunskapsbedömning av NO-ämnet förutsätter detta, liksom att detta ger möjlighet att anknyta den nationella kunskapsbedömningen till TIMSS. Eftersom det dock inte kan förutsättas att en sådan anknytning är realiserbar diskuteras några alternativa uppläggningar.

Ramverket för TIMSS 2007 (Mullis, Martin, Ruddock, O’Sullivan, Arora, & Erberber, 2005) omfattar matematik och naturvetenskap för åk 4 och åk 8. För enkelhets skull behandlas här dock endast åk 8.

För matematik i åk 8 anger ramverket fyra innehållsdomäner: tal (”number”), algebra, geometri, och data och slump, där de två första domänerna skall omfatta 30 % vardera, och de två senare 20 % vardera. Tre kognitiva domäner (”processer”) preciseras: kunna (”knowing”) (35 %), tillämpa (”applying”) (40 %), och resonera (”reasoning”) (25 %). För naturvetenskap specificerar ramverket fyra domäner: biologi (35 %), kemi (20 %), fysik (25 %) och ”earth science” (20 %). Samma kognitiva domäner gäller som för matematik, men med något annorlunda viktning: kunna (30 %), tillämpa (35 %), och resonera (35 %).

För matematik används 14 uppgiftsblock, som vardera innehåller ungefär 10 – 15 uppgifter, eller totalt ca 180 uppgifter. Varje block tar 22,5 minuter, vilket innebär att den totala testtiden för matematik uppgår till 315 minuter. Även för naturvetenskap används 14 uppgiftsblock med 22,5 minuters testtid per block. Sammantaget består sålunda TIMSS i åk 8 av 28 block med en total testtid om 10,5 timmar. Hälften av blocken består av nya uppgifter, och hälften av blocken av uppgifter som användes i TIMSS 2003, vilket gör det möjligt att studera förändring över tid.

De 28 blocken sätts samman till 14 häften, vart och ett bestående av fyra block, två från matematik och två från naturvetenskap. Varje block finns sålunda i två häften, vilket utgör grunden för länkning av alla uppgifterna till en och samma skala. Den effektiva testtiden uppgår till 90 minuter per häfte, vilka genomförs i två pass med en paus. Efter ytterligare en paus besvarar eleverna ett frågeformulär (30 minuter).

Den minimala urvalsstorlek som anges är 4500 elever per land, vilket innebär att minst ca 640 elever bjuds varje uppgift. Det måste dock betonas att detta är en mycket grov bestämning av den minsta urvalsstorleken, som inte tar hänsyn till vare sig urvalsdesign eller beräkningar av den statistiska styrkan att upptäcka förändringar.

Denna grunddesign kan vara en god utgångspunkt för att beskriva både nivå och förändring i kunskaper i matematik och naturvetenskap. Samtidigt står det klart att detta inte ger ett tillräckligt underlag för en kunskapsbedömning som den kan förväntas föreslås bli utformad i svenska ramverk för matematik och naturvetenskap. Som ett exempel kan nämnas att i TIMSS saknas området Teknik, medan däremot ”earth science” ingår, vilket närmast motsvarar vissa av de naturvetenskapligt inriktade delarna av geografiämnet. Mer noggranna jämförelser mellan ramverken för TIMSS och motsvarande svenska ramverk skulle säkerligen peka ut ytterligare punkter där det finns skillnader.

Ett sätt att hantera detta är att konstruera kompletterande uppgifter, som avser Teknik och andra underrepresenterade områden. Om dessa skall inkluderas i skalningen organiseras de tillkommande uppgifterna lämpligen i block med samma egenskaper som de som ingår i TIMSS. Det är då också lämpligt att, om möjligt, skapa ett jämnt antal block inom varje område, varvid det ena publiceras efter den första omgången, och det andra bevaras hemligt för trendmätning vid den andra omgången. Antag, exempelvis, att ett 50-tal kompletterande uppgifter behövs. Detta motsvarar fyra block, vilka kan sättas samman till ett häfte, med en testtid om 90 minuter. Dessa extrahäften kan slumpmässigt fördelas

till vissa eller alla de elever som deltar i TIMSS, vilket innebär att testtiden förlängs från 90 minuter till 180 minuter. Eftersom de 14 häftena fördelas slumpmässigt inom klassrum är det mest praktiskt att extrahäftet fördelas slumpmässigt över klassrum om det inte skall bjudas till samtliga elever. Denna procedur medför att det etableras direkta eller indirekta länkar mellan samtliga uppgifter, varför det inte bör utgöra något problem att genomföra en IRT-skalning som även inkluderar de kompletterande uppgifterna. Genom att denna skalning bygger på flera uppgifter ökas också reliabiliteten i de individuella estimaten.

Om de kompletterande uppgifterna inte skall inkluderas i skalningen tillsammans med TIMSS-uppgifterna kan en något enklare administrationsprocedur användas, eftersom länkar mellan uppgifterna då inte behöver etableras. En möjlighet är exempelvis att slumpvis fördela de fyra blocken till de elever som deltar i TIMSS, varvid testtiden förlängs från 90 till 112,5 minuter. Med denna procedur skapas inga länkar mellan de kompletterande uppgifter som ingår i olika block, men däremot skapas underlag för att undersöka samvariation mellan TIMSS-uppgifter å ena sidan och de kompletterande uppgifterna å den andra.

För den händelse det inte är möjligt att direkt anknyta till TIMSS-undersökningen är det ändå rimligt att använda denna som ett förebildligt exempel på hur en kunskapsbedömning som konstrueras från grunden kan läggas upp. Den exakta utformningen måste givetvis grundas i omsorgsfullt konstruerade ramverk för matematik och naturvetenskap, varför det inte är möjligt att här föreslå en konkret design. Det är dock rimligt att anta att ungefär den mängd uppgifter som används i TIMSS kommer att krävas även i en uppläggning som helt utgår från svenska förhållanden. Inom NU-03 användes exempelvis en total testtid om tre timmar för biologi, fysik och kemi (Andersson et al., 2004), vilket kan jämföras med de dryga fem timmar som används för naturvetenskap i TIMSS. Som redan fastslagits bör det också vara en utgångspunkt i den skalorienterade ansatsen att de uppgiftstyper som används i TIMSS är de dominerande.

Det är dock inte rimligt att förvänta sig att en kunskapsbedömning som är begränsad till dessa typer av uppgifter och undersökningsinstrument skall kunna fånga upp hela bredden och djupet av kunskaper och färdigheter ens inom områden som matematik och naturvetenskap, och än mindre inom andra fält som bild och slöjd.

Inom NU-92 och NU-03 användes exempelvis relativt omfattande gruppuppgifter, där en grupp elever under 120 minuter, uppdelade på ett eller två tillfällen, löste antingen uppgiften att göra ritningar för ett tält eller att planera en avslutningsfest. Gruppuppgiften gavs till delurval av de elever som ingick i matematikutvärderingen. En sådan uppgift låter sig inte enkelt infogas i ett matrissamplingssystem och det är överhuvudtaget knappast meningsfullt att låta resultaten från gruppuppgiften ingå i skalbestämningen. Det finns givetvis en lång rad exempel på komplexa uppgifter av olika omfattning som inte på ett naturligt sätt låter sig inomordnas under den skalorienterade ansatsen.

Ett alternativ är då att använda sig av de komplexa uppgifterna i en uppgiftsorienterad ansats. Detta innebär att resultatredovisningen fokuserar på en mer ingående beskrivning

och analys, ofta av kvalitativt slag, av genomförande och utfall på de enskilda uppgifterna. Dessa är heller normalt inte länkade till övriga uppgifter som bjuds vid ett visst tillfälle. Däremot används samma uppgift vid på varandra följande bedömnings-tillfällen, för att det skall vara möjligt att studera förändring över tid. Detta är i princip den design som används inom NEMP, även om eleverna där deltar i flera uppgifter med en sammantagen tidsinsats om fyra timmar.

Förslagsvis genomförs sålunda de komplexa uppgifter som inte låter sig inordnas i den skalorienterade ansatsen under den uppgiftsorienterad ansats, med en uppläggning som liknar den som används i NEMP. Det är dock lämpligt att dessa uppgifter genomförs av de elever som deltar i den skalorienterade ansatsen, eftersom resultat på dessa delar, liksom givetvis även bakgrundsvariablerna, är av värde vid analysen av de olika uppgifterna. Av praktiska skäl bör dock de komplexa uppgifterna inte randomiseras inom klasser, utan det torde ofta vara lämpligare att alla elever inom en klass gör samma uppgifter. De komplexa uppgifterna kan lämpligen föras samman i grupper på så sätt att den sammanlagda tidsåtgången blir ungefär den samma. Detta är dock inget absolut krav, och om någon uppgift skulle vara extra tidsödande, som gruppuppgiften i matematik, är detta inget hinder. Uppgiftsgruppen fördelas sedan slumpmässigt över ett antal klasser.

I tabellen nedan ges i konkretiserad form ett exempel. Den skalorienterade delen av kunskapsbedömningen antas omfatta 14 häften, vilket vart och ett kräver 90 minuter. Varje häfte besvaras av 450 elever, vilket medför en total undersökningsgrupp om 6300 elever. Alla elever besvarar också en enkät med frågor kring elevens bakgrund och syn på skolan och undervisningen. I TIMSS kräver elevenkäten 30 minuter och 30 minuter har beräknats för en kompletterande elevenkät (se vidare sid 111 nedan). Olika grupper av klasser inom det totala urvalet genomför sedan olika aktiviteter. En grupp, här med en beräknad storlek om 1800 elever, besvarar ett kompletteringshäfte med frågor som också skall ingå i skalningen. Detta är givetvis endast aktuellt om kunskapsbedömningen grundas på en internationell undersökning. Andra grupper av klasser genomför komplexa uppgifter. I exemplet antas att det finns tre grupper av komplexa uppgifter, vilka tar olika tid i anspråk, och för var och en av dessa beräknas ett urval om 1500 elever.

Grupp	N	Aktivitet 1	Aktivitet 2	Aktivitet 3
1	1800	Häfte 1-14, 90 min	Bakgrunds- och kontextenkät, 60 min	Kompletterings häfte, 90 min
2	1500	Häfte 1-14, 90 min	Bakgrunds- och kontextenkät, 60 min	Komplexa uppgifter 1, 90 min
3	1500	Häfte 1-14, 90 min	Bakgrunds- och kontextenkät, 60 min	Komplexa uppgifter 2, 120 min
4	1500	Häfte 1-14, 90 min	Bakgrunds- och kontextenkät, 60 min	Komplexa uppgifter 3, 60 min

Denna design kan användas som en grundläggande modell för kunskapsbedömningar inom alla områden, även om den på alla punkter givetvis måste anpassas efter de krav som ramverket för det aktuella ämnesområdet anger. Antalet häften och mängden tid för varje häfte kan sålunda förväntas variera kraftigt över olika ämnen och ämnesområden, liksom behovet av tid för de komplexa uppgifterna. Då kunskapsbedömningen avser enskilda ämnen snarare än en grupp av ämnen är det rimligt att antalet uppgifter är lägre, och i vissa fall är det måhända tillfyllest med ett häfte. Inom vissa områden är det också rimligt att förvänta sig en större tyngdpunkt på den uppgiftsorienterade ansatsen än på den skalorienterade ansatsen.

Fördjupad analys och förklaring

Även om det primära syftet med det nationella kunskapsbedömningssystemet är att beskriva utvecklingen av kunskaps- och färdigheter på nationell nivå visar de internationella erfarenheterna (se kapitel 5, sid 65) att det är nödvändigt att också ha en möjlighet till fördjupad analys och tolkning av de observerade förändringarna. Skälet till detta är att de olika intressenter som tar del av resultaten också förväntar sig förklaringar, liksom att beredvilligheten hos olika grupper att producera mer eller mindre ad hoc betonade förklaringar är stor. Även om svårigheterna att nå fram till entydiga slutsatser om kasualrelationer är stora, kan ändå ett empiriskt underlag ge bättre förutsättningar för en informerad diskussion, inte minst genom att det ger möjlighet att utesluta vissa förklaringar.

Även i detta sammanhang framträder betydande skillnader mellan den skalorienterade och den uppgiftsorienterade ansatsen. Den senare kan ge möjlighet till kvalitativa analyser av arten av förändring över tid, vilka kan utgöra värdefulla tolkningsunderlag i jakten på förklaringar. Samtidigt ger den skalorienterade ansatsens högre grad av aggregering och precision bättre möjligheter att med statistisk metodik undersöka effekter av determinanter på olika nivåer inom och utom utbildningssystemet (t ex resurser i form av lärartäthet och lärarkompetens, demografiska faktorer, hemmiljön, och utformning av undervisningen).

Det måste dock understrykas att tvärsnittsdata är långtifrån lämpade som underlag för kausalanalyser, och vidare är data från matrissamplingsdesigner och komplexa urvalsmodeller svåra att analysera. Som diskuteras mer utförligt i kapitel 5 (se sid 67) är huvudproblemet att det i tvärsnittsdata är svårt att kontrollera för selektionseffekter, vilket i sin tur beror på att mått på tidigare prestationer inte finns tillgängliga.

Registerdata för analysändamål

Nationell kunskapsbedömning med longitudinell uppläggning skulle ur analyssynpunkt vara att föredra. Svårigheterna att över tid följa enskilda elever gör emellertid att ett sådant angreppssätt är uteslutet, i synnerhet som det är naturligt att i detta sammanhang använda skola som en urvalsenhet. Om det är möjligt att skapa det slags registerbaserade nyckeldatabaser som nämndes i samband med diskussionen om optimering av

urvalsdesignen (se sid 104) skulle det dock vara möjligt att skapa en slags ”virtuell” longitudinell design. Förutsättningen är då att det i register, som helst omfattar samtliga elever, finns information om kunskaper och färdigheter (t ex resultat på nationella prov i åk 5 eller 9, eller betyg) som antingen i tiden föregår eller kommer efter den nationella kunskapsbedömningen. Samkörning av informationen i registren med den information som samlats in inom den nationella kunskapsbedömningen ger då möjlighet att använda registerinformationen antingen som kontrollvariabel då faktorer som har betydelse för resultaten i den nationella kunskapsbedömningen analyseras, eller att använda de senare resultaten som kontrollvariabler och registerinformation (exempelvis de nationella proven i åk 9) som beroendevariabler.

Som redan påpekats ger registerkopplingen också möjlighet till enkel åtkomst av andra för analysen betydelsefulla variabler, som exempelvis elevbakgrund och kontextuella variabler på skolnivå.

Även om denna metod att i efterhand skapa en longitudinell design inte är fullt ut optimal därför att de mått på kunskaper och färdigheter som finns tillgängliga i registren inte är helt jämförbara med den information som samlas in inom den nationella kunskaps-bedömningen ger den på ett mycket kostnadseffektivt sätt förbättrade möjligheter till fördjupade analyser. Det bör därför vara en strävan att söka skapa sådana nyckeldatabaser för analysändamål.

Enkätbaserade insamlingsmetoder

Antingen registerinformation kan nyttjas eller ej kommer enkäter till elever, lärare och skolledare att vara den viktigaste källan till information om olika tänkbara förklarings-faktorer. Information insamlad via enkäter utgör också ett viktigt underlag för att beskriva undervisningens inriktning och uppläggning, även då det inte finns ett uttalat syfte att använda denna information i förklarande syfte. Med hjälp av enkäter insamlas också information om viktiga utfall, som exempelvis elevers egna bedömningar av kunskaper och färdigheter, attityder till olika ämnen och ämnesdelar, och motivation att lära.

De multipla syftena och de oklara och ofta outtalade informationsbehoven gör det till en utomordentligt grannlaga uppgift att konstruera sådana instrument så att de fångar upp de betydelsefulla variablerna, utan att vara överlastade med irrelevanta frågor. Det faktum att många av de aspekter som enkäterna syftar att belysa är ämnesövergripande gör det också viktigt att noggrant planera uppläggning och utformning av enkäterna.

De omständigheter som pekats på ovan gör det omöjligt att i detta läge precisera vilka frågor och frågegrupper som skall inkluderas i enkäterna. Utformningen av innehållet i enkäterna bör istället primärt bygga på två typer av underlag: det ramverk som utvecklas för kunskapsbedömningen, och resultat från analyser av tidigare använda enkäter.

I utvecklingen av ramverket för kunskapsbedömningen ingår att producera ett dokument som beskriver vilka bakgrunds- och kontextvariabler som skall samlas in via enkäter (se sid 100). Som exempel kan nämnas att ramverket för TIMSS 2007 inkluderar ett omfattande ”contextual framework”, med följande huvudkomponenter: läroplan (t ex utvecklings- och beslutsmodell, omfattning och innehåll, organisation av skolsystemet, uppföljning och utvärdering, läroplansmaterial), skolor (demografi, skolorganistion, skolors mål, skolledarens roll, resurser för att stödja undervisningen inom matematik och naturvetenskap, utrustning, skolklimat, föräldramedverkan, lärarrekrytering och utvärdering av lärare), lärarna (utbildning, arbetsuppgifter, fortbildning, och erfarenhet), klassrumsaktiviteter (undervisningsinnehåll, klasstorlek, undervisningstid, undervisningsaktiviteter, läxor, dator- och Internetanvändning, och användning av miniräknare), och eleverna (hembakgrund, attityder, motivation, självbild). Till kontextvariablerna hör sålunda elevernas attityder till ämnet, självbedömning, och motivation, och som tidigare påpekats utgör detta samtidigt viktiga resultat- och förklaringsvariabler.

Den andra typen av underlag för utveckling av enkäter grundas i analyser av tidigare använda enkäter. Medan stor uppmärksamhet ägnas åt mätegenskaperna hos de kognitiva frågorna i kunskapsbedömningar belyses mer sällan reliabilitet och validitet i enkätfrågor. Om dessa frågor skall användas i analyssammanhang är det dock minst lika viktigt att belysa dessa frågors mätegenskaper, och att söka optimera dessa, bland annat genom att sätta samman resultat på flera frågor till sammanfattande mått (se t ex Hansen, Rosén & Gustafsson, i tryck). NU-03 erbjuder ett mycket rikt material av enkätfrågor som väl lämpar sig för denna typ av analyser.

Framförallt måste dock valet av variabler att inkludera grundas i mer ingående, teori-baserade, överväganden om vilka mekanismer och processer som är viktiga. Dessa överväganden måste också influeras av resultaten från tidigare analyser. Endast genom aktiva försök att genom forskning hitta förklaringar till de observerade resultaten är det sålunda möjligt att förbättra informationsunderlaget.

För att enkäterna skall vara användbara för analysändamål är det viktigt att de är systematiskt konstruerade över olika ämnen och över tid. Detta gör det nödvändigt att utformningen av enkäterna sker på ett samordnat sätt över de olika ämnena och ämnesgrupperna. I synnerhet gäller detta bakgrundsvariabler och ämnesoberoende frågor. Det är också viktigt att även de frågor som avser ett visst ämne (t ex ämnets betydelse, arbetssätt i ämnet) ges en standardiserad utformning så att det är möjligt att göra jämförelser mellan ämnen.

Speciella problem uppstår om de internationella studierna skall användas som delkomponenter i den nationella kunskapsbedömningen, eftersom dessa i allmänhet omfattar enkäter till elever, lärare och skolledare. Det finns sålunda en risk att dessa grupper behöver besvara två enkäter vardera, en för den internationella studien, och en för den nationella kunskapsbedömningen. I synnerhet som dessa enkäter skulle komma att omfatta delvis likartade frågor framstår inte detta som en lämplig uppläggning.

Ett möjligt sätt att lösa detta problem är att de enkäter som ingår i den internationella studien används i oförändrat skick, och att kompletterande enkäter utformas med de frågor som behövs för att motsvara den nationella kunskapsbedömningens krav. Denna lösning gör det möjligt att delta fullt ut i den internationella studien, men eftersom det kan vara svårt att använda frågorna i de internationella enkäterna för de nationella syftena kan det krävas omfattande kompletteringsenkäter, med ty åtföljande tidskrav. En annan möjlighet är att endast använda enkäter utformade för den nationella kunskaps-bedömningen, och avstå från de enkäter som ingår i den internationella studien. Detta skulle innebära ett partiellt deltagande i den internationella studien, med utnyttjande endast av kunskapsproven. Eftersom enkätinformationen i de internationella studierna även används i estimeringen, och enkätvariablerna har en central roll i den internationella rapporteringen skulle detta innebära att Sverige inte skulle kunna ingå i redovisningen av de internationella resultaten, vilket givetvis är en stor nackdel. Denna fråga kräver ytterligare utredning innan det är möjligt att göra ett ställningstagande.

Slutsatser

Det är ofrånkomligt att ett nationellt kunskapsbedömningssystem också måste kunna ge underlag för förklaringar och tolkningar av de trender som observeras. Såväl uppgifts- som skalorienterade ansatser torde vara av värde i detta sammanhang, även om den skalorienterade ansatsen framstår som den mest användbara för analytiska ändamål.

Ett enkelt och kostnadseffektivt sätt att förstärka analysmöjligheterna vore att skapa s k nyckeldatabaser genom sammanläggning av registerinformation med information insamlad inom den nationella kunskapsbedömningen.

Stor omsorg måste också ägnas åt insamling av information kring bakgrundsvariabler och tänkbara förklaringsvariabler. Detta görs vid utvecklingen av ramverket, och en annan viktig grund är resultat från försök att i olika typer av forskning hitta förklaringar till de observerade utfallen.

Databasuppbyggnad

Enligt uppdraget från Skolverket ingår

> Att diskutera hur databasuppbyggnad och rapporteringsrutiner skall utformas så att de på ett flexibelt sätt medger fortsatta fördjupade analyser av insamlade resultatdata.

Det är en utomordentligt komplex uppgift att bygga upp en databas som omfattar ett stort antal variabler från många informationskällor. Komplexiteten ökas ytterligare av att det måste vara möjligt att göra jämförelser såväl över tid som över ämnen. Ett ytterligare krav är att data måste vara enkelt tillgängliga för sekundäranalys, vilket ställer mycket stora krav på en komplett och lättillgänglig dokumentation.

För att det skall vara möjligt att motsvara dessa krav är det nödvändigt med strikt standardisering och gemensamma rutiner, för vilka omfattande datorstöd behövs. Inom ramen för arbetet med de internationella studierna har sådana rutiner och programvaror utvecklats under en lång tidsrymd. Eftersom kraven på databaserna i den nationella kunskapsbedömningen i stort sammanfaller med kraven på databaserna i de internationella studierna bör de erfarenheter som gjorts i de senare tas till vara. Ett sätt att göra detta kan vara ett anlita exempelvis IEAs ”Data Processing Center” (DPC) i Hamburg för arbetet med att skapa databaserna. DPC är en omfattande organisation som utöver arbete med IEAs internationella studier åtar sig uppdrag att konstruera databaser för externa uppdragsgivare. En annan möjlighet skulle kunna vara att anlita DPC som konsult för att ge stöd vid uppbyggandet av en motsvarande svensk organisation med uppgift att tillgodose databasbehoven för den nationella kunskapsbedömningen. Denna fråga behöver utredas närmare innan det är möjligt att ge ett mer konkret förslag.

Former för redovisning av resultat

Den information som ett nationellt kunskapsbedömningssystem skapar är av intresse för ett stort antal grupper: allmänheten, elever, lärare, skolledare, beslutsfattare på lokal och nationell nivå, läroplansutvecklare, läroboksförfattare, och lärarutbildare, för att endast nämna några. Arten och utformningen av den information som olika grupper behöver varierar dock.

I vissa sammanhang är det nödvändigt att få en överblick över huvuddragen i skolsystemets utveckling, och inte minst beslutsfattare på nationell nivå, liksom den skolintresserade allmänheten, är primärt intresserad av denna typ av översiktlig och aggregerad information. Den skalorienterade ansatsen kan ge sådan information, medan möjligheterna för den uppgiftsorienterade ansatsen att göra det är betydligt mer begränsade.

I andra sammanhang behövs en betydligt mer detaljerad information, exempelvis kring utvecklingen inom olika delar av ett ämnesområde, eller kring förändringar i elevers sätt att angripa olika problem. Grupper som på olika sätt är aktivt involverade i skolans arbete, eller i olika former av utvecklingsarbete, är i behov av sådan mer innehålls- och verksamhetsnära information. Den uppgiftsorienterade ansatsen kan ge denna typ av information, medan däremot den skalorienterade ansatsens möjligheter är mer begränsad därvidlag.

De påtagligt olika informationsbehoven pekar på att det är nödvändigt att på ett optimalt sätt utnyttja informationen både från den skalorienterade och den uppgiftsorienterade ansatsen.

Skalorienterad rapportering

Som framgår av diskussionen i kapitel 6 har mycket uppmärksamhet ägnats åt att utveckla modeller för standardsbaserad rapportering inte minst inom NAEP (se sid 81). Huvudsyftet med detta har varit att på ett enkelt, konkret och tillgängligt sätt

kommunicera huvudresultaten även till grupper som inte har erfarenhet av tolkning av statistisk information. En nackdel är dock att den standardsbaserade redovisning i mycket sammanfaller med utformningen av det mål- och kunskapsbaserade betygssystemet, vilket för grundskolans del är kopplat till mål och kriterier för åk 9. Det vore olyckligt om det nationella kunskapsbedömningssystemet skulle uppfattas som ett parallellt betygssystem, vilket är ett skäl för att avstå från att redovisa resultat med hjälp av standardsbaserade redovisningsmodeller. Ytterligare ett skäl för detta är att det fortfarande finns tveksamhet hos många forskare om i vilken utsträckning de metoder som utvecklats för att fastställa gränserna mellan de olika nivåerna i standardsbaserade redovisningsmodeller generar rimliga och pålitliga resultat (se t ex Pellegrino et al., 1999). Kostnaden för att tillämpa dessa metoder är också hög, eftersom de kräver omfattande utbildnings- och arbetsinsatser för stora grupper av bedömare. Förslaget är sålunda att det nationella kunskapsbedömningssystemet inte utnyttjar standardsbaserade redovisningsmodeller.

Det enklaste alternativet är att redovisa resultaten från den skalorienterade ansatsen i form av poäng och poängförändringar på en standardiserad skala. En möjlighet är exempelvis att fastställa medelvärdet till 500 och standardavvikelsen till 100 för det år som är startår för en trendlinje, och att sedan förändringar uttrycks i denna skala. En fördel med en sådan skala är att poängförändringarna är lätta att översätta till effekt-storlek, eftersom detta mått i allmänhet är uttryckt i standardavvikelseenheter. Konkret innebär detta att en förändring med effektstorleken 0,10, vilken enligt tidigare diskussion (se sid 101) framstår som intressant att kunna upptäcka, motsvarar en förändring om 10 poäng. Även om det innebär en viss pedagogisk utmaning framstår det inte som en omöjlighet att förse presentationerna av resultatförändringar med tolkningsanvisningar som gör siffrorna tolkbara och meningsfulla även för grupper som inte har omfattande erfarenhet av användning av kvantitativa metoder.

Tidsramarna för rapporteringen av de skalorienterade resultaten kan vara ett problem, och enligt uppdragsbeskrivningen från Skolverket skall den första redovisningen av resultaten ske inom tolv månader. Inom NAEP har det varit ett närmast ständigt problem att rapporteringen av de initiala resultaten varit långsam på grund av de komplexa analyserna, och ofta har de första resultaten kommit först efter ca två år. Målsättningen nu är att den initiala rapporteringen av de huvudsakliga resultaten skall göras inom 6 månader efter datainsamling. Även för de internationella studierna dröjer det i allmänhet ca två år mellan datainsamling och rapportering. Anledningen till tidsutdräkten är framförallt att det tar mycket lång tid att kontrollera all information och foga samman alla observationer till en komplett databas. Eftersom de betingade MML-skattningarna kräver att både resultaten på kunskapsbedömningsuppgifterna och bakgrundsvariablerna finns tillgängliga måste hela databasen vara konstruerad innan skattningarna kan genomföras (Forsyth et al., 1996). Uppgiften försvåras givetvis av att det både i NAEP och i de internationella studierna handlar om mycket stora mängder data från ett stort antal delstater eller länder, liksom att man i allt större utsträckning använder komplexa uppgifter som kräver omfattande bedömningsinsatser.

Om den statistiska modellen inte utnyttjar betingning på bakgrundsvariabler förenklas uppgiften eftersom endast kunskapsbedömningsuppgifterna då behöver läggas in i databasen innan resultaten kan räknas fram. Som redan påpekats finns det även av analysmässiga skäl anledning att försöka undvika betingning på bakgrundsvariabler, och ett sätt att göra detta är att säkerställa att tillräcklig mängd information samlas in för att få tillräckligt hög reliabilitet på individnivå. Detta innebär att mängden tid som varje elev deltar i kunskapsproven behöver utsträcks från ca 50 minuter till åtminstone det dubbla. De designer som i exemplifierande syfte skisserats ovan (se sid 106) uppfyller detta krav, och ett förslag är därför att skattningarna genomförs utan betingning på bakgrundsvariabler.

Eftersom det nationella kunskapsbedömningssystemet endast är en svensk angelägenhet finns ingen anledning att invänta de internationella resultaten i de fall då samordning sker med de internationella studierna. Detta innebär att skattningen kan ske så snart som de svenska kunskapsproven rättats/bedömts, registrerats, kontrollerats och fogats samman till en databas på individnivå.

Under antagande att bakgrundsvariabler inte används i MML-skattningen och att skattningarna endast avser svenska elevers data torde det vara fullt genomförbart att erhålla skattningarna mellan sex till tolv månader efter genomförd datainsamling. En förutsättning för detta är dock att bedömningen av elevproduktion inskränker sig till ett begränsat antal kortsvarsuppgifter.

Uppgiftsorienterad rapportering

Den uppgiftsorienterade ansatsen har inte samma centrala roll i bestämningen av förändringstender som den skalorienterade ansatsen, även om också denna kan förväntas bidra med mycket värdefull information om förändringarnas omfattning och art. Informationen från den uppgiftsorienterade ansatsen är dock av större intresse för målgrupper som arbetar nära innehållet och skolverksamheten än för politiker och beslutsfattare. Kravet på snabb redovisning torde vara större från de senare kategorierna än från de förra, medan däremot kraven på djup i analysen och omfattningen av redovisningen torde vara större från de förra kategorierna.

Dessa omständigheter gör det naturligt att hävda att kravet på rapportering inom tolv månader skall begränsas till att omfatta resultaten från den skalbaserade ansatsen. Resultat från den uppgiftsorienterade ansatsen kan givetvis föreligga inom tolv månader, men det framstår inte som ändamålsenligt att ha detta som ett generellt krav. En ytterligare anledning till detta är att den tid som krävs för bedömning och analys av det material som eleverna producerat i mycket stor utsträckning varierar över uppgifter, vilket gör det svårt att göra generella bedömningar av tidsåtgång.

Inte heller vad gäller utformningen av den uppgiftsorienterade redovisningen torde det vara möjligt att ställa upp några generella regler eller rekommendationer, eftersom detta i hög grad är beroende av uppgifternas art, och analysens ambitionsnivå. De fördjupade

forskarrapporterna inom NU-03 kan i många fall ses som exempel på uppgiftsorienterad rapportering, och här föreligger en stor variation över de olika ämnena.

Webbaserad rapportering

Inom NAEP och NEMP görs den initiala rapporteringen i form av tryckta resultatrapporter av relativt begränsad omfattning. Inom NAEP har man också med stor framgång börjat använda webbaserad rapportering. Den stora fördel som erbjuds av dessa är att användarna själva har möjlighet att med enkla metoder genomföra även komplexa analyser. Dessa analyser kan avse såväl resultat på enskilda uppgifter som resultat på olika skalor. Det är givetvis ofta av stort intresse att analysera de olika uppgifter som ingår i de olika skalorna, vilket innebär att såväl de uppgifter som ingår i den uppgiftsorienterade ansatsen som de som ingår i den skalorienterade ansatsen bör göras tillgängliga för analys.

Mot bakgrund av de framgångar som det webbaserade analys- och rapporteringssystemet för NAEP har fått framstår det som lämpligt att utreda kostnaderna för att utveckla och implementera ett sådant system för det nationella kunskapsbedömningssystemet. Ett välfungerande sådant system skulle eventuellt göra det möjligt att begränsa omfattningen av den tryckta resultatredovisningen till att omfatta en kortfattad och lättillgänglig presentation av de initiala resultaten.

Fördjupade analyser

Såväl den uppgiftsorienterade ansatsen som den skalorienterade ansatsen kräver mer omfattande och fördjupade analyser, där syftet både är att ge mer ingående beskrivningar av resultaten och att hitta förklaringar. Inom NAEP genomförs ett mycket stort antal sådana studier, både inom ramen för på förhand planerade specialstudier, och i form av sekundäranalyser som genomförs av forskare. I syfte att ge utrymme för och stimulera till sådana sekundäranalyser ställer NCES forskningsmedel och stipendier till förfogande att söka för intresserade forskare. Även inom NEMP genomförs fördjupade analyser som presenteras i en speciell rapportserie (”probe reports”).

Som framgått av diskussionen ovan är ett aktivt analysarbete en nödvändig förutsättning både för ett framgångsrikt sökande efter förklaringar till de observerade trenderna, och för den fortsatta utvecklingen av instrumenten. Det är därför nödvändigt att den nationella kunskapsbedömningens data används i olika forskningssammanhang, och att resultat presenteras i olika vetenskapliga fora. Medel för detta måste finnas tillgängliga.

Lokal utvärdering

Enligt uppdraget från Skolverket är ”Det … önskvärt att skolor/skolhuvudmän ges möjlighet till att på eget initiativ genomföra lokala utvärderingar vilkas resultat kan relateras till de nationella.”

Inom NAEP har det över årtiondena skett en förändring på så sätt att allt lägre nivåer inom utbildningssystemet har inkluderats, från den nationella över delstatsnivån till större skoldistrikt. De större skoldistrikt (t ex New York, Los Angeles, och Chicago) omfattar dock i de många fall fler elever än vad som finns i Sverige, och ambitionen är inte att göra uttalanden som avser skolnivå. Det finns åtminstone två skäl för detta. Det ena är att det skulle vara mycket komplext och kostsamt att genomföra en kunskapsbedömning med de angreppssätt som här beskrivs så att tillräcklig säkerhet skulle uppnås på de lägsta aggregationsnivåerna, och detta gäller både den skal- och den uppgiftsorienterade ansatsen. Det andra skälet är att resultaten skulle vara svåra att tolka och då i synnerhet på skolnivå. Anledningen till detta är att de val som gjorts vad gäller innehåll och metoder i undervisningen inte behöver sammanfalla så väl med kunskapsbedömningens utformning eftersom denna måste sträva efter en bred täckning. Det framstår därför inte som lämpligt att genomföra de lokala utvärderingarna på så sätt att den nationella kunskapsbedömningen utvidgas till att omfatta skolor i fler kommuner.

Ett annat sätt att utnyttja metoder och resultat från den nationella kunskapsbedömningen i lokalt utvärderingsarbete vore att utnyttja de uppgifter som frisläpps efter användning. Inom NAEP och NEMP publiceras mellan 30 % och 50 % av uppgifterna efter varje omgång, medan resten bevaras hemliga för trendmätning. De publicerade uppgifterna har en viktig uppgift att fylla i samband med publiceringen av resultaten för att illustrera och konkretisera dessa. Men uppgifterna görs också tillgängliga för lokal användning. Inom NAEP har ett särskilt webbaserat system utvecklats för utsökning och distribution av uppgifter för lokal användning, vilket visat sig vara mycket framgångsrikt.

Förslagsvis införs även i Sverige stöd för att på lokal nivå använda det frisläppta uppgiftsmaterialet, vilket skulle medföra en lång rad fördelar. De uppgifter som utvecklas för den nationella kunskapsbedömningen kommer att vara omsorgsfullt konstruerade och noga utprövade, och vara försedda med omfattande anvisningar för bedömning av elevsvar. Användningen av uppgifterna kommer också att generera ett rikligt empiriskt material i form av elevsvar och bedömningar av dessa, liksom givetvis även ett statistiskt jämförelsematerial. Uppgifterna och åtminstone vissa delar av det empiriska materialet kan läggas in i en databas där det är fritt tillgängligt för lokal användning.

Databasen skulle omfatta såväl de uppgifter som ingår i den skalorienterade ansatsen, som de som ingår den uppgiftsorienterade ansatsen. Användningen skulle dock huvudsakligen vara uppgiftsorienterad och baseras på ett urval av uppgifterna. Av de skäl som anfördes ovan är det i allmänhet inte aktuellt att försöka bestämma skalpoäng, eftersom detta kräver användning av ett stort antal uppgifter som administreras enligt samma procedurer som i den nationella bedömningen. Även de uppgifter som utvecklats för användning i den skalorienterade ansatsen kan dock givetvis användas på ett uppgiftsorienterat sätt, och i många fall omfattar även dessa uppgifter elevproduktion som kan vara värdefull i formativa och diagnostiska sammanhang.

Inom Skolverket finns redan nu en Provbank, som framförallt innehåller provmaterial för summativ bedömning på gymnasienivå i de ämnen där det inte finns nationella prov.

Denna Provbank skulle med fördel kunna utvidgas till att omfatta även de frisläppta uppgifterna från den nationella kunskapsbedömningen.

Ledningsformer och förankring

De internationella erfarenheterna pekar på att det är viktigt att olika intressegrupper får inflytande över det nationella kunskapsbedömningssystemet, och att systemet är förankrat hos de betydelsefulla aktörerna. Inte minst är detta viktigt därför att det är oundgängligen nödvändigt att systemet är stabilt över lång tid.

Som framgår av beskrivningen av NAEPs utveckling har frågan om den breda representationen i formuleringen av policy varit central, vilket lett till en mycket komplex ledningsstruktur med NAGB och NCES som de viktigaste organen. Även i arbetet med utvecklingen av ramverk har strävan varit att uppnå en bred representation av olika synsätt och kompetenser. Den breda förankringen av NAEP är sannolikt en förklaring till att detta program överlevt så länge, och utvecklats så väl. Även NEMP arbetar aktivt med en stor och brett sammansatt referensgrupp.

Ett förslag är att det tidigt utses en referensgrupp för det nationella kunskapsbedömnings-systemet med representation av bland andra kommuner, lärarfackliga organisationer, och elev- och föräldraorganisationer.

Bilaga 1. Utvärderingar genomförda inom NAEP från 1969 till 2004.

YEAR	NATIONAL	STATE	LONG–TERM TREND
1969–70	citizenship science writing	State assessments began in 1990	science[1]
1970–71	literature reading		reading[1]
1971–72	music social studies		
1972–73	mathematics science		mathematics[1] science[1]
1973–74	career/occupational development writing		
1974–75	art index of basic skills reading		reading[1]
1975–76	citizenship/ social studies mathematics[2]		citizenship/ social studies[1]
1976–77	basic life skills[2] science		science[1]
1977–78	consumer skills[2] mathematics		citizenship/ social studies[1]
1978–79	art music writing		
1979–80	reading literature art		reading[1]
1981–82[3]	mathematics science citizenship social studies		mathematics[1] science[1]
1984	reading writing		reading writing
1986	computer competence U.S. history[2] literature[2] mathematics science reading		mathematics science reading[4]
1988	Civics, document literacy[2] geography[2] U.S. history		civics[1] science reading

	reading writing		writing
1990	mathematics science reading	mathematics (8)[5]	mathematics science reading writing
1992	mathematics reading writing	mathematics (4, 8)[5] reading (4)[5]	mathematics science reading writing
1994	geography U.S. history reading	reading (4)[5]	mathematics science reading writing
1996	mathematics science	mathematics (4, 8) science (8)	reading writing[6] mathematics science
1997	arts (8)		
1998	reading writing civics	reading (4, 8) writing (8)	
1999			reading mathematics science[7]
2000	mathematics science reading (4)	mathematics (4, 8) science (4, 8)	
2001	U.S. history geography		
2002	reading writing	reading (4, 8) writing (4, 8)	
2003	reading (4, 8) mathematics (4, 8)	reading (4, 8) mathematics (4, 8)	
2004	foreign language (12)[8] (postponed)		reading mathematics

[1] This assessment appears in reports as part of long-term trend. Note that the civics assessment in 1988 is the third point in trend with citizenship/social studies in 1981–82 and in 1975–76. There are no points on the trend line for writing before 1984.

[2] This was a small, special-interest assessment administered to limited national samples at specific grades or ages and was not part of a main assessment. Note that this chart includes only assessments administered to in-school samples; not shown are several special NAEP assessments of adults.

[3] Explanation of format for year column: Before 1984, the main NAEP assessments were administered in fall of one year through spring of the next. Beginning with 1984, the main NAEP was administered after the new year in winter, although the assessments to measure long-term trend continued with their traditional administration in fall, winter, and spring. Because the main assessment is the largest component of NAEP, beginning with 1984 we have listed its administration year rather than the two years over which trend continued to be administered. Note also that the state component is administered at essentially the same time as the main NAEP.

[4] The 1986 long-term trend reading assessment is not included on the trend line in reports because the results for this assessment were unusual. Further information on this reading anomaly is available in *Beaton and Zwick (1990)*.

[5] State assessments in 1990–94 were referred to as trial state assessments (TSA).

[6] After 1996, long-term trend in writing was no longer reported because of technical reasons having to do with the relatively small number of writing prompts. See remarks by the commissioner.

[7] After the 1999 long-term trend in science, it was determined that technical studies are required to enable necessary changes to the design and revisions to the item pool in order to maintain the long-term trend in this subject. New items will be developed and field tested for use in future assessments. For that reason, the science long-term trend assessment will not be given in 2004. For more information, see the NAGB policy (114K Microsoft Word document) on long-term trend assessments.

[8] The National Assessment Governing Board postponed the foreign language assessment at their March 6, 2004 meeting.

Bilaga 2.

Planerade utvärderingar inom NAEP 2005 - 2017

YEAR	NATIONAL	STATE	LONG–TERM TREND
2005	reading mathematics[1] science high school transcript study	reading (4, 8) mathematics (4, 8)[1] science (4, 8)	
2006	U.S. history civics economics (12)[1]		
2007	reading (4, 8) mathematics (4, 8) writing (8, 12)	reading (4, 8) mathematics (4, 8) writing (8)	
2008	arts (8)		reading mathematics
2009	reading[1] mathematics science[1] high school transcript study	reading (4, 8)[1] mathematics (4, 8) science (4, 8)[1]	
2010	U.S. history civics geography[1]		
2011	reading (4, 8) mathematics (4, 8) writing[1]	reading (4, 8) mathematics (4, 8) writing (4, 8)[1]	
2012	world history (12)[1] foreign language (12)[1] probe: technological literacy (special study)[1]		reading mathematics
2013	reading mathematics science high school transcript study	reading (4, 8) mathematics (4, 8) science (4, 8)	
2014	U.S. history[1] civics[1] geography		
2015	reading (4, 8) mathematics (4, 8) writing	reading (4, 8) mathematics (4, 8) writing (4, 8)	

YEAR	NATIONAL	STATE	LONG–TERM TREND
2016	arts (8)		reading mathematics
2017	reading mathematics science high school transcript study	reading (4, 8) mathematics (4, 8) science (4, 8)	

[1] Updated or new framework is planned for implementation for this subject. Framework for foreign language approved by Governing Board May 2000; updates to mathematics framework approved November 2001; economics framework approved August 2002. In the case of subjects for which frameworks are already adopted (i.e., reading, writing, mathematics, science, the arts, U.S. history, geography, and civics), the Board will decide whether a new or updated framework is needed for this assessment year.
Note: Grades tested are 4, 8, and 12 unless otherwise indicated, except that long-term trend assessments sample students at ages 9, 13, and 17 and are conducted in reading and mathematics.

Referenser

Adams, R. J. (2002). Scaling PISA cognitive data. I R. J. Adams & M. Wu (Eds.). *PISA 2000 Technical Report*, s 99 - 108. Paris: OECD Publications.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Alexander, L., & James, H. T. (Eds.) (1987). *Improving the assessment of student achievement: The nation's report card.* Washington, D. C.: National Academy of Education.

Barron, S. I., & Koretz, D. M. (1996). An evaluation of the National Assessment of Educational Progress trend estimates for racial ethnic subgroups. *Educational Assessment, 3*(3), 209-248.

Barron, S. I. (2000). Difficulties associated with secondary analysis of NAEP data. I N. S. Raju, J. W. Pellegrino, M. W. Betenthal, K. J. Mitchell, & L. R. Jones (Eds.) *Grading the Nation's Report Card. Research from the Evaluation of NAEP,* (s 172-194). Washington, D. C.: National Academy Press.

Barton, P. (2002). Perspectives on Background Questions in The National Assessment of Educational Progress. Report to the National Assessment

Bay, L., Chen, L., Hanson, B. A., Happel, J., Kolen, M. J., Miller, T., Pommerich, M., Sconing, J., Wang, T., & Welch, C. (1997). *ACT's NAEP REDESIGN PROJECT: Assessment design is the key to useful and stable assessment results* (Working Paper No. 97-30). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Beaton, A. E., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191–204.

Beaton, A. E., & Johnson, E. (2004). Emerging technical innovations in NAEP. I L. V. Jones & I. Olkin (Eds.) *The Nations Report Card. Evolution and Perspectives,* (pp. 449-466). Bloomington, Indiana: Phi Delta Kappan.

Beaton, A. E., & Zwick, R. (1990, February). *Disentangling the NAEP 1985-86 reading anomaly* (NAEP Rep. No. 17-TR-21). Princeton, NJ: Educational Testing Service.

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4-11+16.

Braun, H. (2005, 14 april) Personlig kommunikation.

Brown, W. (2000). Reporting NAEP by Achievement Levels: An Analysis of Policy and External Reviews. I Bourque, M. L., & Byrd, S. (Eds.) *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements,* (s 13-39). Washington, D. C.: National Assessment Governing Board.

Campbell, J. R., Hombo, C. M., Mazzeo, J. (2000). *NAEP 1999 Trends in Academic Progress. Three Decades of Student Performance.* U. S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, NCES 2000-469. Washington, D. C.

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001 (2nd ed.).* Chestnut Hill, MA: Boston College.

Champagne, A., & Pearson, D. P. (2003). Subject domain: What is being measured? I *NAEP Validity Studies: An Agenda for NAEP Validity Research*. National Center for Education Statistics, Working Paper no. 2003-07.
Cohen, J. (2002). AM statistical software. American Institutes for Research. http://am.air.org/default.asp
Elley, W. B. (1994) (Ed). *The IEA study of reading literacy: Achievement and instruction in thirty-two school system*. Oxford, England: Pergamon.
Finn, C. E., Jr. (2004, February 2). Education in urban America. *Hoover Institution Weekly Essays.* http://www-hoover.stanford.edu/pubaffairs/we/2004/finn02.html.
Flockton, L. (1999). *School-wide assessment. National Education Monitoring Project*. Wellington: New Zealand Council for Educational Research.
Flockton, L., & Crooks, T. (1999). New Zealand's National Education Monitoring Project: The First Four Year cycle, 1995 - 1998. Paper presented at the combined annual conference for 1999 of the New Zealand Association for Research in Education and the Australian Association for Research in Education, Melbourne, Australia
Gilmore, A. M. (1999). The NEMP Experience. Paper presented at the combined annual conference for 1999 of the New Zealand Association for Research in Education and the Australian Association for Research in Education, Melbourne, Australia
Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *The trial state assessment: Prospects and realities. The third report of the National Academy of Education panel on the evaluation of the NAEP trial state assessment: 1992 trial state assessment.* NY: The National Academy of Education.
Grissmer, D., Beaton, A. E., & Hedges, L. (2003). Estimating trends from NAEP scores: Rationale and Research Directions. I *NAEP Validity Studies: An Agenda for NAEP Validity Research*. National Center for Education Statistics, Working Paper no. 2003-07.
Gustafsson, J.-E. (1997). Measurement characteristics of the IEA Reading Literacy scales for 9-10 year-olds at country and individual levels. *Journal of Educational Measurement*, 34(3), 233-251.
Gustafsson, J.-E., & Rosén, M. (i tryck). The Dimensional Structure of Reading Assessment Tasks in the IEA Reading Literacy Study 1991 and the Progress in International Reading Literacy Study 2001. *Educational Research and Evalutation*.
Hansen, K. J., Rosén, M., & Gustafsson, J.-E. (in press). Measures of self-reported reading resources, attitudes and activities based on latent variable modeling. *International Journal of Research and Method in Education*.
Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., et al. (2000). A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' *Grading the Nation's Report Card*. *Educational Measurement: Issues and Practice, 19*(2), 5–14.
Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990 to 1998. I Bourque, M. L., & Byrd, S. (Eds.) *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements*, (s 131-156). Washington, D. C.: National Assessment Governing Board.

Hambleton, R. K., & Slater, S. (1996, April). *Are NAEP executive summary reports understandable to policymakers and educators?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
Hedges, L. V., & Vevea, J. L. (2003). *NAEP Validity Studies: A Study of Equating in NAEP*. National Center for Education Statistics, Working Paper no. 2003-13.
Härnqvist, K. (1975). The International Study of Educational Achievement. *Review of Research in Education*, *3*, 84-109.
Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher*, *25*(7), 15-22.
Jones, L. V. (1999). The assessment of student achievement: The hundred years war. Invited address, AERA Annual Convention, Montreal, Canada.
Kaplan, D. (1995). The impact of BIB-spiraling induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, *20*(1), 69-82.
Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: Educational Testing Service.
Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey* (NCES 93-275). Washington, DC: U.S Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.
Kolen, M. J. (2000). Issues in phasing out Trend NAEP. I N. S. Raju, J. W. Pellegrino, M. W. Betenthal, K. J. Mitchell, & L. R. Jones (Eds.) *Grading the Nation's Report Card. Research from the Evaluation of NAEP*, (s 132-151). Washington, D. C.: National Academy Press.
Lazer, S. (2004). Innovations in instrumentation and dissemination. I L. V. Jones & I. Olkin (Eds.) *The Nations Report Card. Evolution and Perspectives*, (pp. 469-488). Bloomington, Indiana: Phi Delta Kappan.
Linn, R. L. (2004). The influence of external evaluations. I L. V. Jones & I. Olkin (Eds.) *The Nations Report Card. Evolution and Perspectives*, (pp. 291-308). Bloomington, Indiana: Phi Delta Kappan.
Linn, R. L., & Baker, E. L. (1996). Assessing the validity of the National Assessment of Educational Progress: NAEP Technical Review Panel White Paper. Center for the Study of Evaluation, CSE Technical Report 416. Los Angeles: UCLA.
Lord, F. M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement*, *22*, 259-267.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Kennedy, A. M. (2003). *Trends in children's reading literacy achievement 1991-2001*. Boston: Lynch School of Education.
McLaughlin, D. H. (2001). *Exclusions and accommodations affect state NAEP gain statistics: Mathematics, 1996 to 2000.* Report to the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.
Mehrens, W.A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of the Joint Conference on Standard Setting for Large Scale*

Assessments, Volume II (pp. 221-263). Washington, DC: National Assessment Governing Board.
Messick, S. (1989). Validity. I Linn, R. (Ed.) *Educational Measurement* (3rd ed). Washington: National Council of Measurement in Education.
Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81–91.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–61.
Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.
Mislevy, R. J., & Sheehan, K. J. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (Rep. No 15-TR-20, pp. 293–360). Princeton, NJ: Educational Testing Service.
Mislevy, R. J., & Sheehan, K. J. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661–679.
Mullis, I. W., Campbell, J. R., & Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States: Data from the National and Trial State Assessments*. NCES Report No. 23-ST06; September. Washington, D. C.: U. S. Department of Education.
Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C., Arora, A. & Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center.
Muthén, B., Khoo, S., & Goff, G. N. (1994). Multidimensional description of subgroup differences in mathematics achievement data from the 1992 National Assessment of Educational Progress. Technial report. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
NAGB (2005). Reading framework for the 2009 National Assessment of Educational Progress. Pre-publication edition. http://www.nagb.org/
Pellegrino, J. W., Jones, L. R., and Mitchell, K. J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.
Raudenbush, S. W., Bryk, A. S., Cheong, Y. F,, & Congdon, R. (2004). *HLM6: Hierarchical Linear and Nonlinear Modeling.* Chicago, Ill: Scientific Software International.
Reckase, M. D. (2000) A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. I Bourque, M. L., & Byrd, S. (Eds.) *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements,* (s 43-69). Washington, D. C.: National Assessment Governing Board.
Reckase, M. D. (2004, June). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts completed by ACT.* ACT, Inc.: Iowa City, IA.

Rust, K. F., & Johnson, E. G. (1992) Sampling and weighting in the national assessment. *Journal of Educational Statistics, 17,* 111–129.
Skolverket (2004). Internationella studier under 40 år – Svenska resultat och erfarenheter. Skolverket: Stockholm
Stone, J. C. (1990). Book review of G. G., Madaus & D. Stufflebeam (Eds.). Educational Evaluation: Classic Works of Ralph W. Tyler. *Educational Policy and Evaluation Analysis*, 12(1), 102-106.
Stuffelbeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural effort to set achievement levels on the National Assessment of Educational Progress.* Kalamazoo: Western Michigan University, The Evaluation Center.
Söderberg, S. (2005). Reseberättelse från studiebesök på CITO, Arnhem, Nederländerna 2005-12-06. Skolverket.
Tyler, R. W. (1970). National assessment: A history and sociology. *School and Society, 98*, 471-477.
U. S. Department of Education, National Center for Education Statistics (1999). N. Horkay (Ed.). *The NAEP Guide*. NCES report No. 2000-456.
U. S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations.* GAO/PEMD-93-12. Washington, DC: U. S. Government Printing Office.
Van Lent, G., & Bakker, S. (2004). Monitoring educational achievement. 'Where practical constraints hit theoretical requirements – looking for acceptable shades of gray'. Paper presented at the annual meeting of the International Association for Educational Assessment, Philadelphia, USA.
Vinovskis, M.A. (1998). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board*. Washington, DC: National Assessment Governing Board.
Wainer, H. (1994a). *On the academic achievement of New Jersey's public school children. I. Fourth & eighth grade mathematics in 1992.* (TR94-3) Princeton, NJ: Educational Testing Service.
Yang-Hansen, K., Rosén, M., & Gustafsson (i tryck). Measures of self-reported reading resources, attitudes and activities based on latent variable modeling. *International Journal of Research Methods in Education*.
Yepes-Baraya, M. (1996). A cognitive study based on the National Assessment of Educational Progress (NAEP) Science Assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, April 1996.
Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293-308.