Rätt mätt på prov

Om validering av bedömningar i skolan

Peter Nyström

Pedagogiska institutionen, Umeå universitet
Nr 71

Pedagogiska institutionen
Umeå universitet
Avhandling 2004

Tryckt på Umeå universitets tryckeri
Februari 2004

ISSN 0281-6768
ISBN 91-7305-609-X

Nyström, Peter. Rätt mätt på prov. Om validering av bedömningar i skolan [Validation of educational assessments]. Dissertation at the Faculty of Social Sciences, Umeå University, Sweden, 2004. ISBN 91-7305-609-X, ISSN 0281-6768.

Abstract

The thesis focuses on validation as a process whereby the quality of educational assessments can be evaluated. On the basis of Messick's (1989) validity concept, which takes consequences as well as inferences of assessment results into account, a model for validation is described. An argument made is that assessment purposes, epistemology, and curricular goals are necessary starting points for validation. Furthermore, it is argued that several of the criteria for good assessment proposed as alternatives to validity and reliability can be used together with Messick's framework for validity and make it useful in practice. Four empirical studies, each presented in a separate paper, supply examples of the application of this framework. In paper I it is claimed that reliability is a problem inherent in all educational assessments, and it is given a meaningful interpretation through the concept of classification accuracy. The article examines how the classification accuracy of a Swedish national test in mathematics is affected by changes in the test that are motivated by other validity concerns. The study is based on the results of 1,201 students participating in a Swedish national test in mathematics. The results indicate that there can be a significant trade-off between reliability and other aspects of validity. Paper II presents a study of attainment differences in mathematics related to the grouping of students into different study programmes in the Swedish upper-secondary school. For a sample of 403 students, results from two consecutive national tests in mathematics were compared, one at the end of compulsory school and the other at the end of the first mathematics course in upper-secondary school. The results indicate that attainment is, on average, positively affected if students are grouped together with higher-achieving peers. This effect appears to be strongest for low-achievers. Paper III addresses questions concerning how perceptions of competence are formed in school mathematics. The study is based on a questionnaire distributed to a sample of Swedish upper secondary school students (n = 550). The results indicate that an inner feeling of understanding is most highly valued by the students as an indicator of competence, more than external cues like test results and teacher feed-back. Male and female students seem to value cues to perceived competence in similar ways. In paper IV an

interview study with six mathematics teachers in upper-secondary school is presented, aiming at exploring the variation in what teachers at their particular school say about ability grouping. In summary, what these teachers express concerning the advantages and disadvantages of ability grouping varies a great deal. There are, however, some things they more or less agree on. They have a basically positive view of the way in which ability grouping has been practised at their school. They identify problems with ability grouping for low-ability groups, primarily related to the mixing of students who have trouble learning with those who primarily have motivational problems. The list of advantages and possibilities is definitely longer for groups of high-achieving students compared to groups of low-achievers.

Förord

Avhandlingens titel ”Rätt mätt på prov” kan tolkas på många sätt, och jag vill inledningsvis hjälpa läsaren genom att ge tre förslag till tolkningar.

En tolkning av rubriken skulle kunna vara att jag som skriver detta skulle vara ganska trött på att diskutera prov och andra bedömningar i skolan. Det ligger en viss sanning i det utifrån att en avhandling kräver ett ganska intensivt arbete med texten och problemen, vilket innebär att man till slut bara önskar få stänga ordbehandlingsprogrammet och göra något annat. Trots detta kan jag dock säga att mitt intresse för frågor om bedömning i skolan är större än någonsin.

En annan tolkning av rubriken skulle kunna vara att någon annan än jag skulle vara rätt så mätt på prov. Skolverkets utvärderingar har visat på att dagens skola i många stycken är en ”provskola”. Detta är inget särdrag hos den svenska skolan, utan liknande och ännu mer långtgående trender rapporteras från flera håll. Det finns anledning att tro att såväl lärare som elever tycker att det ibland blir för mycket. En hel del ”sanningar” kring prov och bedömning behöver nog omprövas, gällande till exempel former för bedömning, vem som kan och ska bedöma, samt bedömningars plats i förhållande till undervisning och skolans mål. Mina egna erfarenheter av att granska elevers svar på olika uppgifter som getts i bedömningssyfte har också ibland gett en insikt i den vånda och känsla av misslyckande som ganska många elever måste uppleva. Bland annat för de studerandes skull är det angeläget att bedömningar görs med höga krav på kvalitet när det gäller slutsatserna som lärare och andra kan dra, och att de även görs med olika konsekvenser av bedömningar och bedömningssituationer i åtanke. Det är också angeläget att prov och andra former för bedömning betraktas som en integrerad del i undervisning och lärande.

En tredje tolkning indikerar att det skulle finnas rätt sätt att mäta måluppfyllelse i skolan, och därmed också sätt som är fel. Så svartvitt och enkelt är det naturligtvis inte. Denna tolkning ligger i alla fall närmast det som avhandlingen handlar om, nämligen hur vi kan värdera förtjänster och brister hos alla de bedömningar av individers och gruppers måluppfyllelse som görs i skolsammanhang.

Det finns många personer som betytt mycket för mig under avhandlingsarbetet, och både gjort denna avhandling möjlig och bidragit till att göra den till vad den är. Ett särskilt tack vill jag rikta till alla skolledare, lärare och

elever som jag mött och som välvilligt och intresserat ställt upp i mina projekt.

Jag vänder mig också till alla mina vänner på Enheten för pedagogiska mätningar, i synnerhet kollegorna inom projektet Nationella prov och provbank. Tack för era bidrag till en positiv arbetsmiljö som gjort att jag alltid tyckt om att komma till jobbet, trots en resväg på 14 mil enkel väg. Ett stort tack till min handledare Widar Henriksson, för ditt oändliga tålamod när det gäller att läsa och ge värdefulla kommentarer, och för din ständiga uppmuntran. Jag vill också särskilt tacka Jan-Olof Lindström, som efter ett oväntat möte på en pizzeria uppmanade mig att söka jobb på Pedagogiska mätningar, och som varit en stor inspirationskälla genom sitt kunnande och sin aldrig sinande energi. För ett utmärkt samarbete, och för alla diskussioner som hjälpt mig att skärpa argumentationen och tänkandet riktar jag också ett varmt tack till Torulf Palm. Jag vill även tacka Susanne Alger och Bitte Wallin för hjälp med språkgranskning av mina engelska texter, och Lotta Jarl för hjälp med slutredigeringen. Även Birgitta Törnkvist, Statistiska institutionen, är väl värd ett omnämnande för den hjälp jag fått med statistiska spörsmål.

Jag vill även vända mig till alla jag mött på Pedagogiska institutionen, i samband med kurser med mera. Jag har varit ”distansstuderande” i och med att jag haft min arbetsplats på annat håll, men har alltid känt mig välkommen till Beteendevetarhuset, och alltid blivit positivt bemött. Ett särskilt tack till seminariegruppen, ingen nämnd och ingen glömd. Jag har uppskattat gemenskapen och lärt mig oerhört mycket av våra diskussioner.

Framförallt, och ytterst, vill jag tacka min familj: Maria, Viktor, Arvid, Sara och Alma. Utan ert stöd hade det inte gått.

Skellefteå, en vacker söndag i februari 2004
Peter Nyström

Rätt mätt på prov
Om validering av bedömningar i skolan

Artiklarna som avhandlingen omfattar är

I. Nyström, P. (under tryckning).Reliability of educational assessments – The case of classification accuracy. *Scandinavian Journal of Educational Research*.

II. Nyström, P. (2003). National tests as a means of evaluating effects of streaming in Swedish upper secondary school mathematics. Insänd för publicering.

III. Nyström, P. (2003). Students' cues to perceived competence in mathematics. Insänd för publicering.

IV. Nyström, P. (2003). Lika barn leka bäst? En gymnasielärardiskurs om nivågruppering i matematik. *Pedagogisk Forskning i Sverige 8*(4), 225-246.

Alla hänvisningar till dessa artiklar kommer att använda sig av ovanstående numrering.

Innehållsförteckning

Avhandlingens syfte och disposition

Övergripande syfte

Det övergripande syftet med denna avhandling är att utveckla en utgångspunkt för värdering av kvalitet i samband med bedömningar i skolan. Detta görs utifrån ett allomfattande validitetsbegrepp, och den modell för validering som förordas bygger på explicitgörandet av den syn på kunskap, läroplan och bedömningens syfte som ligger till grund för utformningen av bedömningssituationer och genomförandet av bedömningen. De bedömningar som diskuteras här avgränsas till skolan, det vill säga alla formella utbildningssammanhang som regleras av en läroplan.

De fyra artiklar som ingår i avhandlingen behandlar olika problem och forskningsfrågor, men de innehåller också gemensamma teman. Det mest framträdande gemensamma temat handlar om validering av bedömningar. Jag avser att lyfta fram de olika studiernas relevans för valideringsfrågorna där så är möjligt, samt utveckla de utgångspunkter för validering som ligger till grund för det empiriska arbete som presenteras i artiklarna.

Disposition

Avhandlingen disponeras så att det närmast följande avsnittet definierar och problematiserar några begrepp som är viktiga och flitigt använda genom hela avhandlingen. I nästa avsnitt tas validitetsbegreppets definition och användning upp. Där tolkas och diskuteras det validitetsbegrepp som Messick (1989) introducerat, och kombineras med andra kriterier på kvalitet i bedömningar. Därefter kommer tre avsnitt som problematiserar och fördjupar de tre områden som jag menar är centrala utgångspunkter för valideringen, nämligen bedömningens syfte, läroplanens mål samt kunskapssynen. Så kommer en kort sammanfattning som ska peka på hur de olika delarna hänger ihop. I påföljande avsnitt sammanfattas de artiklar som ingår i avhandlingen, och sedan presenteras några vetenskapsteoretiska utgångspunkter för det empiriska arbete som presenteras i artiklarna. Så presenteras några övergripande reflektioner, varefter avhandlingen avslutas med en engelsk sammanfattning och en referenslista.

Några viktiga begrepp

Inledningsvis är det angeläget att definiera och problematisera några begrepp och termer som är relevanta för det forskningsområde som avhandlingen berör. Avsikten är i första hand inte att göra en översikt över alla olika betydelser som de använda begreppen kan ha, utan det är framförallt ett försök att definiera några begrepp som används i denna avhandling. Eftersom begreppsapparaten inte är särskilt väl utvecklad på svenska så kan detta avsnitt också utgöra ett bidrag till en svensk terminologi på området.

’Bedömning i skolan’ står i fokus i denna avhandling, och detta uttryck representerar för det första ett forskningsfält. Fältet kännetecknas av att det studerar bedömningspraktiken i utbildningssammanhang i alla dess former, och att detta görs utifrån ett brett samhällsvetenskapligt perspektiv. Teoribildningar och erfarenheter från den mätinriktade forskningen kring prov är en viktig del i detta fält, men bedömning är något mer än prov och perspektivet inom fältet ’bedömningar i skolan’ är vidare än inom psykometrin. Fältet kan likställas med det som Gipps (1994) kallar ”educational assessment”, och som enligt henne skiljer sig från de tidigare dominerande begreppen psykometri respektive ”educational measurement”. Ett nytt begrepp som beskrivning av fältet motiveras, enligt Gipps, med att utvecklingen från åttiotalet och framåt har satt läraren och lärandet i centrum för bedömningsfrågorna, till skillnad från tidigare synsätt i England och USA där läraren enbart förväntades administrera test som utformades av någon annan.

> … educational measurement has now become called more generally educational assessment; this is largely because ’measurement’ implies a precise quantification, which is not what the educational assessment paradigm is concerned with. (Gipps, 1994, sid. 10).

Gipps, och även Shepard (2000), menar att det har skett ett paradigmskifte under 1990-talet som innebär att tidigare dominerande psykometriska synsätt och utgångspunkter ersatts av mer pedagogiska och sociologiska perspektiv på bedömning. Det finns anledning att ifrågasätta om denna beskrivning stämmer för svenska förhållanden. Svenska lärare har aldrig varit reducerade till förmedlare av test som utformas av någon annan och psykometriska utgångspunkter har knappast varit särskilt framträdande när det gäller bedömningar i skolan. Oavsett om uttrycket ’bedömningar i skolan’ signalerar ett paradigmskifte eller inte så representerar det en allsidig belysning av problem relaterade till bedömning.

För det andra representerar termen ’bedömningar i skolan’ alla de aktiviteter som används i utbildningssammanhang för att värdera individers och gruppers måluppfyllelse. Termen bedömning ska ses som en svensk motsvarighet till ”assessment”. Ingen av dessa termer representerar något entydigt begrepp, men vi kan urskilja såväl en snävare som en vidare betydelse som båda har.

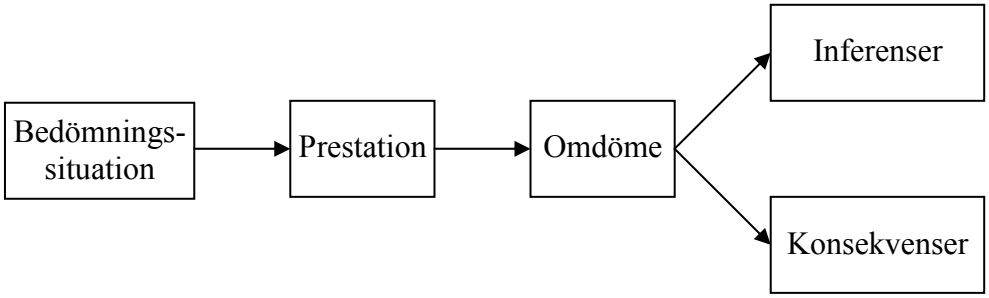
Den snävare tolkningen av bedömning (och assessment) syftar på en värdering av en prestation i förhållande till en bedömningsanvisning av något slag. Till exempel görs ofta en bedömning av studerandes prestationer på skriftliga prov utifrån en rättningsmall och resultatet av bedömningen redovisas som en provpoäng. Detta kräver ingen tolkning av vad provpoängen betyder. En sådan bedömning gäller isolerat vad den studerande presterat i förhållande till de uppgifter som givits och är ingen generalisering till mer övergripande kompetenser.

I den vidare tolkningen betraktas bedömning (och assessment) som en process som innehåller flera steg. Airaisian (1994, sid. 5) ger prov på en sådan vidare syn på bedömning i sin definition: ”The process of collecting, synthesizing, and interpreting information to aid in decision making is called *assessment*.” Här betecknar alltså bedömning hela processen från observationer till tolkningar och även beslut. Wiggins (1993) går tillbaka till latinet för att tolka vad assessment betyder. Ordet kommer från det latinska verbet assidere, som betyder ’sitta med’.

> Such a “sitting with” suggests that the assessor has an obligation to go the extra mile in determining what the student knows and can do. The assessor must be more tactful, respectful, and responsive than the giver of tests – more like “a mother and a manager”, in British researcher John Raven’s phrase, “than an imperious judge”. (Wiggins, 1993, sid. 14)

Bedömning utgår alltså från en strävan att skapa bästa möjliga förutsättningar för alla att visa vad de kan.

Som ett underlag för att identifiera och problematisera bedömningar i skolan kan hela bedömningsprocessen schematiskt beskrivas med modellen i Figur 1. En värdering av kvalitet hos bedömningar i skolan handlar ytterst om att granska alla stegen i bedömningsprocessen.



Figur 1 Modell för bedömningsprocessen.

I modellen syftar *bedömningssituation* på en uppgift, men också på en vidare kontext som innefattar olika ramar för prestation och bedömning, till exempel villkor för vilka hjälpmedel som får användas och hur arbetet ska redovisas. Uppgiftens grad av struktur kan variera, alltifrån en öppen, ostrukturerad, uppgift som kanske är ett sammanhang som ”i flykten” definieras som en bedömningssituation av de inblandade, till mycket strukturerade uppgifter, där s.k. flervalsuppgifter är den mest utpräglade typen. Bedömningssituationer kan vara formella, det vill säga utformade specifikt för olika bedömningssyften, och informella. Informella bedömningssituationer kan vara sådana som sker i klassrummet utan att det är klart uttalat att bedömning sker. Termen prov används här som beteckning för alla formella bedömningssituationer.

Prestation är ett gensvar på bedömningssituationen. Prestationen kan handla om såväl produkt som process, dvs. bedömningar kan värdera kvaliteten hos såväl slutresultat av ett arbete som hur arbetet fortskrider under tiden det pågår. Betoningen av att det är en prestation som bedöms bygger på att det endast är det som uttrycks som kan bedömas. Vi kan aldrig värdera kunskap och kompetens hos en individ som inte uttrycker den på något sätt.

Den del av bedömningsprocessen som i Figur 1 kallas *omdöme* är primärt en värdering av prestationen i förhållande till bedömningssituationen, det vill säga hur väl bedömningsobjektet har hanterat de uppgifter och ramar som bedömningssituationen innehåller. Detta direkta omdöme kan bokföras kvantitativt, i form av poäng, eller kvalitativt, i form av skriftliga omdömen. I allmänhet stannar dock inte bedömningen vid ett omdöme om den direkta prestationen på de uppgifter som givits, utan det görs tolkningar och generaliseringar på grundval av prestationen. Till exempel kan omdömet om elevers lösning till en viss skrivuppgift ligga till grund för slutsatser om elever-

nas skrivkompetens. I enlighet med Messick (1989) kan ett primärt omdöme, till exempel en provpoäng, sägas ha såväl inferenser som konsekvenser.

Inferenser och *konsekvenser* syftar på de medvetna och omedvetna generaliseringar som sker på grundval av omdömet. En prestation är alltid avgränsad i tid, rum och kontext, och den utförs inte alltid under optimala förhållanden. Slutsatser om en individs måluppfyllelse bygger därför alltid på inferenser från en bedömning. Ett exempel på en inferens är att studerande anses uppfylla målen för ett kursmoment utifrån prestationen på det begränsade urval av uppgifter som ges i ett prov. Ett annat exempel är diagnoser, som kan leda till värderingen att en elev inte på ett tillfredsställande sätt behärskar vissa kunskapsområden, och som även kan ligga till grund för handlingar i form av lämpligt stöd och hjälp för elevens lärande. Konsekvenser syftar till exempel på den påverkan som bedömningssituationer och omdömen kan ha för den bedömdes självbild och syn på ämnet. I nästa kapitel kommer inferenser och konsekvenser att diskuteras mer ingående i samband med validitetsbegreppet.

Även om det ofta är så att den som bedömer och den som blir bedömd är olika personer, så vill jag betona att bedömningar i skolan även innefattar situationer där gränser mellan objekt och subjekt suddas ut. Det mest framträdande exemplet på detta är självvärdering. Även för sådana bedömningsformer är det angeläget att värdera kvaliteten i bedömningarna. Ett annat exempel är att makten över villkoren för bedömning kan variera alltifrån helt lärarstyrda situationer till situationer där eleven är delaktig i såväl val och utformning av uppgiften som förhandling om vilka kriterier som ska gälla för bedömningen.

Validitet

Det finns olika sätt att se på validitet, och begreppet har utvecklats under de mer än 60 år som det varit en central utgångspunkt för utformning och utvärdering av bedömningar (se Wolming, 1998). Jag har valt att använda mig av Messicks (1989) allomfattande validitetsbegrepp därför att det omfattar såväl bedömningars inferenser som deras konsekvenser. I det följande ska jag försöka beskriva min tolkning av Messicks begrepp, men också visa på några andra modeller och bidrag till validering som kopplar ihop Messicks teoretiska begrepp med praktisk validering av bedömningar. Messick uppfattas ofta som svårtolkad och hans teori är inte sällan missförstådd. Jag vill bidra till vidgad förståelse för hur denna validitetsteori kan ge underlag för en värdering av bedömningars kvalitet.

Samuel Messick definierar validitet som

> …an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (Messick, 1989, sid. 13).

Samtidigt som han betonar att validitet är ett sammanhållet begrepp så menar Messick att det kan struktureras genom att man betraktar det från två olika, men ändå sammanlänkade, perspektiv.

Det ena perspektivet beskriver utgångspunkter för värdering av bedömningens inferenser och konsekvenser. Det mest framträdande syftet med bedömningar är att ge belägg för förekomst av och kvalitet hos t.ex. kunskaper. En viktig aspekt av validitet är därför en värdering av kvaliteten i dessa belägg. Är de inferenser som görs på grundval av en prestation rimliga och riktiga? Denna aspekt är vad tidigare definitioner av validitet helt har fokuserat på. Förutom den mer direkta rollen att bidra med belägg för kunskap och kompetens, har bedömningar ett antal mer eller mindre önskvärda och medvetna konsekvenser. De påverkar till exempel vad den som bedöms uppfattar som viktigt och kan även medföra konsekvenser som inte ligger i linje med de belägg som samlas in.

Det andra perspektivet beskriver provets olika funktioner. Messick delar upp dessa i tolkning respektive användning av bedömningens utfall. Med tolkning av utfallet avses den påverkan som detta har på våra uppfattningar och värderingar. Med användning av bedömningens utfall avses de handlingar som görs utifrån resultatet. Genom att kombinera dessa två perspektiv

konstruerar Messick en modell som identifierar de områden som en validering kan och bör omfatta (se Tabell 1). I tabellens rutor anges vad validering inom respektive område huvudsakligen inriktar sig på. Observera att Messick använder termerna prov (test) och provpoäng (test score), men han poängterar att dessa ska ges en generell betydelse som innefattar varje sätt att observera och dokumentera beteenden eller egenskaper.

Tabell 1 Messicks validitetsbegrepp (Messick, 1989, sid. 20).

		Provets resultat och funktion	
		Tolkning av provresultat	*Användning av provresultat*
Utgångspunkter för värdering av provets utformning	*Provets inferenser*	**1** Begreppsvaliditet	**3** Begreppsvaliditet Relevans/nytta
	Provets konsekvenser	Påverkan på värderingar **2**	Sociala konsekvenser **4**

I ruta 1 görs bedömningar om de tolkningar som görs utifrån bedömningen är rimliga. Här spelar begreppsvaliditeten en stor roll. Om vi till exempel vill kunna dra slutsatser om elevers muntliga förmåga i matematik så måste vi fråga oss om bedömningen kan förväntas återspegla en sådan. Möjliggörs en bedömning som indikerar god kommunikativ förmåga, och inte bara en allmän ”mångordighet”? I artikel I argumenterar jag varför reliabilitetsproblematiken i första hand hör hemma i denna ruta. Ruta 2 handlar om bedömningens påverkan på olika värderingar. Hur påverkar till exempel en muntlig del i ett nationellt kursprov lärares och elevers syn på hur viktigt det är med muntlig kommunikation i matematikundervisningen? Det kan också, som diskuteras i artikel III, handla om hur elevers upplevda kompetens påverkas av bedömningar.

I ruta 3 återfinns frågor om den konkreta användningen av ett omdöme är rimlig utifrån den utformning som bedömningen har. I Artikel II diskuteras

möjligheten att använda nationella prov för att utvärdera effekterna av den elevgruppering som skapas av gymnasieskolans uppdelning i program. I ruta 4 frågar vi oss vilka sociala konsekvenser som användningen av omdömet kan ha. Till exempel kan den som bedöms vara kompetent inte uppleva sig som kompetent och därför välja att avbryta sina studier. Ett annat exempel är att bedömningar kan påverka hur studerande bedriver sina studier.

Genom att se till att alla fyra fälten så att säga täcks med validitetsfrågor kan en mångsidig validering säkerställas. Observera att varje tolkning och användning, det vill säga varje syfte med bedömningen, måste valideras för sig.

Det mest karakteristiska i Messicks validitetsbegrepp ligger i att det även innefattar konsekvenserna av bedömningar. Det innebär till exempel att bedömningars påverkan på den upplevda kompetensen, vilket studeras i artikel III, betraktas som en validitetsfråga. Det råder dock delade meningar om sociala konsekvenser ska ses som en del av validitetsbegreppet. Kritikerna hävdar att det visserligen är viktigt att ta hänsyn till sociala konsekvenser av provanvändning, men att detta inte bör blandas ihop med validitet (Mehrens, 2002). Popham (1997) menar att validitetsbegreppet bör användas exklusivt för bedömning av rimligheten i de slutsatser som dras om provtagaren på grundval av provet, det vill säga för värdering av provets inferenser. Messicks validitetsbegrepp innebär en utvidgning till att även omfatta bedömningens konsekvenser som kan motiveras bland annat utifrån att denna aspekt av kvalitet hos bedömningar annars riskerar att tappas bort (se till exempel Shepard, 1997). Detta är också det viktigaste skälet att jag valt att använda mig av Messicks validitetsbegrepp.

Andra kvalitetsmått för bedömningar

I slutet av åttiotalet och början av nittiotalet publicerades ett antal artiklar där de klassiska måtten på bedömningars kvalitet inte ansågs tillräckliga eller adekvata för nya bedömningssituationer, se till exempel Frederiksen och Collins (1989), Haertel (1991), Linn, Baker & Dunbar (1991) och Moss (1992). Det debatterades mellan företrädare för en dominerande testkultur, med i huvudsak uppgifter av flervalstyp (ofta kallade ”objektiva prov”), och förespråkare för så kallad alternativ bedömning. Alternativen benämndes ”performance assessment” och ”authentic assessment”, begrepp som är vanliga i litteraturen men som har en varierande innebörd (Palm, 2001). En tolkning är att i ”performance assessment” sker bedömning mer direkt på de

komplexa kompetenser som undervisningen strävar mot, istället för på isolerade, endimensionella, och relativt små delar som ofta är fallet i samband med ”objektiva” prov.

I jämförelser mellan kvaliteter hos mer traditionella kunskapsmätningar och ”alternativa” bedömningar kan, enligt Linn et al. (1991), argumenten för traditionella mätningar i hög grad vara beroende på utformningen av traditionella valideringsansatser, där reliabilitet, effektivitet och jämförelser mellan bedömningar betonas. De hävdar att andra kvalitetsaspekter på bedömningar kan innebära att ”performance assessments” visar sig överlägsna sina mer ”objektiva” motsvarigheter. Flera sådana listor med kvalitetskriterier för bedömningar har presenterats (Gipps, 1994; Linn et al., 1991; Linn & Herman, 1997). Baker, Linn och Herman (1996) använder sig av fyra utgångspunkter för kvalitetssäkring av prov och andra bedömningssituationer. Dessa är begreppsvaliditet, rättvisa, trovärdighet och användbarhet. Sådana kriterier ska naturligtvis inte föredras därför att de möjligen förändrar styrkan i argumenten för eller emot ”alternativa” bedömningsformer, utan därför att de beskriver viktiga egenskaper och effekter hos bedömningar.

Som ytterligare ett kriterium för värdering av bedömningar tar Linn och Herman (1997) upp samstämmighet mellan bedömningar och läro- och kursplaner. De anser att proven ska byggas upp från grunden efter att målen i läro- och kursplaner är fastställda och då verkligen spegla kunskaper och färdigheter som anges i läro- och kursplaner. Begreppet samstämmighet (alignment) har fått ökad användning vid kvalitetssäkring av prov i USA. Samstämmighet innebär att två eller flera systems komponenter, till exempel bedömningar och läroplan, överensstämmer.

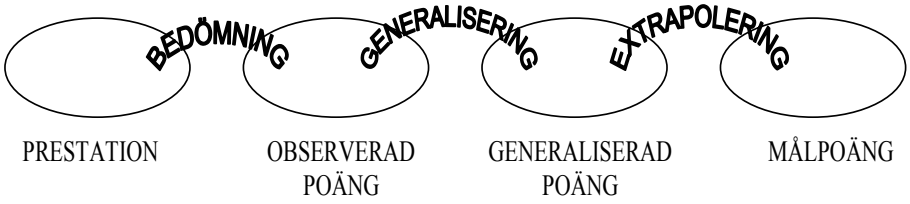
> *Alignment* is the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system towards students learning what they are expected to know and do. (Webb, 1997, sid. 4)

Vad de studerande förväntas veta och kunna syftar närmast på mål som uttrycks i en läroplan.

Hur förhåller sig då dessa kriterier för värdering av bedömningar till Messicks validitetsbegrepp? Jag menar att alla kvalitetsaspekter på prov och bedömningar ryms inom det utvidgade validitetsbegreppet (se även Messick, 1994). För att göra den teoretiska utgångspunkten som beskrivs i Tabell 1 till något användbart behövs meningsfulla frågor och kriterier som kan täcka de olika aspekterna av validitet. I den litteratur som omnämnts ovan finns

många goda exempel på sådana frågor och kriterier. Jag vill även redogöra för ytterligare en modell som kan bidra till konkretisering av validitetsbegreppet.

Bedömningsprocessens steg från prestation till inferenser om måluppfyllelse (se Figur 1) har av Kane, Crooks och Cohens (1999) delats upp i flera steg i en modell som Tindal (2002) kallar "one of the most elegant strategies for addressing the validation process" (p. 6). Modellen beskriver denna process med en broanalogi (se Figur 2).



Figur 2 Modell för bedömning och validering (Kane et al., 1999)

Bron längst till vänster representerar en *bedömning* av prestationens kvalitet som resulterar i en observerad poäng. Författarna talar om poäng, men modellen är lika relevant för andra sätt att bokföra prestationers kvalitet. Nästa bro symboliserar en *generalisering* från prestationen på uppgifterna i en bedömningssituation till alla liknande uppgifter. Det gap som denna bro överbryggar handlar om att prestationen är kontextuell och att kunna lösa en uppgift, med en viss utformning, betyder inte nödvändigtvis att man kan lösa andra uppgifter av liknande slag. Den tredje bron innebär en *extrapolering* från slutsatser om hur en viss uppsättning uppgifter klaras till hur målen för lärandet uppnåtts.

Validering beskrivs som en bedömning av styrkan i dessa broar, och broarnas styrka hänger ihop med hur bedömningssituationerna utformas. För skriftliga prov med en långt driven standardisering och strukturering, som använder sig av flervalsfrågor, är framförallt bron längst till höger problematisk. Om målen kännetecknas av komplexa kompetenser kan det vara mycket svårt att argumentera för hur de enkla uppgiftsformaten som ofta framförallt prövar fakta och grundläggande begreppsförståelse, kan säga något om hur eleven klarar till exempel problemlösning på högre nivå. För bedömningssituationer som bygger på mycket få, stora och komplexa uppgifter är det mit-

tenbron som vållar allra mest bekymmer. Den prestationsnivå som eleverna visar beror på vilken eller vilka uppgifter som eleven arbetar med. I en rad studier har en stark interaktion mellan person och uppgift kunnat påvisas, vilket innebär svårigheter i generaliseringen från en, eller några få, uppgifter till alla motsvarande uppgifter (se t.ex. Brennan & Johnson, 1995). Det betyder att inferenser som baseras på en enda, eller några få uppgifter, riskerar att vara osäkra, vilket kan beskrivas som en del av reliabilitetsproblemet (se artikel I). För sådana bedömningssituationer kan dock bron längst till höger vara mycket stark, även om lärandemålen är komplexa. Eleven får möjlighet att visa komplexa kunskaper och kompetenser eftersom uppgiften simulerar lärandemålen på ett relativt nära sätt.

Messicks teoretiska ramverk lägger grunden för en allsidig och heltäckande validering. Vilka konkreta frågor som bör ställas i förhållande till Messicks validitetsaspekter (de fyra rutorna i Figur 1) avgörs av bedömningssituationens syfte. De ovan nämnda ”alternativa” kvalitetsaspekterna, och den broanalogi som redovisas i Figur 2, ger goda exempel på olika sådana frågor och perspektiv, och vilka som används eller betonas vid en validering avgörs i hög grad av bedömningens syfte. För att valideringen ska vara trovärdig krävs en allsidig belysning, där olika perspektiv lyfts fram och där bedömningens validitet betraktas ur olika aktörers perspektiv. Det är också angeläget att hela bedömningsprocessen beaktas i valideringen (Figur 1) bland annat därför att till exempel bedömningssituationens utformning är relevant för bedömningens inferenser och konsekvenser.

Messick i praktiken

Jag menar alltså att Messicks validitetsbegrepp på ett övergripande plan sammanfattar kvalitetsaspekter för bedömningar på ett strukturerat sätt. Messicks modell bejakar och beskriver på ett meningsfullt sätt också de konflikter eller motsättningar som finns bland validitetsargumenten. Till exempel relationen mellan reliabilitet och samstämmighet (se artikel I). Med ett sådant validitetsbegrepp blir inte reliabiliteten en förutsättning för validiteten (Moss, 1994), vilket gäller för andra tolkningar av validitet. Messicks modell är dock teoretisk, dvs. den ger ingen direkt vägledning till hur validering ska ske konkret. Det är upp till var och en att fylla de olika rutorna med de frågor som är angelägna i förhållande till bedömningens syfte. Hur kan då Messicks modell användas i praktiken? Jag vill försöka peka på möjligheterna genom ett enkelt exempel.

Det är inte ovanligt att lärare säger att eleverna bedöms kontinuerligt. Bedömning är integrerat med klassrumsarbetet och ”bedömer är något jag gör hela tiden”. Detta kan betraktas som ett ambitiöst försök från lärarens sida att göra bedömningen allsidig och att kunna bedöma process och inte bara produkt. Vilka blir slutsatserna om vi skärskådar en sådan praktik utifrån Messicks modell? När det gäller inferenser så finns det förutsättningar för att bedömningen ska bli mångsidig, men samtidigt får man se upp med vad som egentligen bedöms. För informell bedömning är riskerna särskilt stora att man bedömer andra saker än man egentligen avser att bedöma. Detta är en värdering av bedömningens inferenser.

Vilka konsekvenser får det att lärare säger sig bedöma hela tiden? En stor risk är att elever uppfattar det som att man inte får göra fel på lektionerna. De kan bli tysta och inte våga visa upp sina ofullständiga och preliminära förståelser, av rädsla att få en negativ bedömning som kan påverka betyget. Trots att det tycks ”progressivt” att inte låta till exempel ett slutprov vara avgörande för betyget kan en ambition om ständig bedömning vara mycket skadlig för lärandemiljön. I själva verket kan det vara väsentligt att skilja mellan lärande och formativ bedömning å ena sidan och summativ bedömning å den andra (se sidan 13 för en förklaring av begreppen formativ och summativ). Eleverna bör få möjlighet att lära sig utan att hela tiden bli bedömda (för betygssättning) och de bör få ställa hur dumma frågor som helst utan att riskera betyget. Men eleverna ska också veta vid vilken tidpunkt de förväntas ha skaffat sig en god förståelse och uppnått de mål som utbildningen eftersträvar.

Eftersom bedömningens inferenser är direkt kopplade till bedömningens syfte så måste syftet klargöras för att valideringen ska vara möjlig. I följande avsnitt diskuteras olika aspekter av bedömningars syften.

Syften med bedömningar

Enligt mitt sätt att se det syftar bedömningar i skolan ytterst till att främja lärandet, de är med andra ord ett läromedel. Detta är uppenbart när det gäller formativ bedömning, det vill säga bedömning som görs mitt i en undervisningssekvens och har till syfte att ge lärare och elever underlag för fortsatt undervisning och lärande. Black och Wiliam (1998) menar att forskningen mycket konsekvent har visat på positiva effekter av formativ bedömning på elevernas lärande. Men även summativa bedömningar, som syftar till att sammanfatta utfallet av en undervisningssekvens, är ytterst till för lärandets skull. Även statliga initiativ att på nationell nivå bedöma individer och grupper används som underlag för beslut som är tänkta att förbättra lärandet i olika avseenden. En beskrivning av att bedömningar syftar till lärande är dock för generell för att vara intressant och användbar. För att en validering enligt Messick ska vara meningsfull är de nödvändigt med mer preciserade syften och vi ska därför titta litet närmare på hur syften med bedömningar beskrivits i några olika sammanhang.

Det finns väldigt många som har skrivit om de syften som bedömning kan ha, och i de flesta amerikanska och brittiska läroböcker i ämnet finns ett avsnitt om detta, se till exempel Airasian (1994) och Black (1998). I en av de få svenska läroböckerna i ämnet fokuserar Wedman (1988) olika slags prov snarare än olika syften med bedömningar. Han menar att man i skolsammanhang brukar skilja mellan diagnostiska prov, betygsättande prov och allmänt utvärderande prov. Wedman lyfter också fram att man brukar skilja mellan målrelaterade och normrelaterade prov. Han hävdar att det kanske är mer precist att tala om prov med olika syften och användningar, och betonar att dessa indelningar är absoluta.

Detta leder oss in på frågan om en och samma bedömningssituation kan ha flera olika syften och funktioner, vilket också tas upp i artikel II. Gipps (1994) menar, liksom Wedman, att en bedömningssituation inte kan användas för olika syften. Black & Wiliam (1998) har en annan uppfattning. De hävdar att de två funktioner hos bedömningar som brukar kallas formativa respektive summativa inte bara kan utan också bör kombineras i en bedömning. I Gilmore (2002) hänvisas till forskare som anser att ett enskilt instrument inte ska ha mer än en funktion, men där hävdas också att det är fler och fler som anser att provprogram måste fylla flera syften.

Jag menar att bedömningssituationer mycket väl kan ha olika syften, och att det är rimligt att den tid som läggs ned på bedömningsaktiviteter i skolan bör ge så mycket ”återbäring” som möjligt i form av olika funktioner och möjligheter. Det är dock viktigt att varje syfte valideras för sig. Att en bedömning håller hög kvalitet i förhållande till ett visst syfte behöver inte alls betyda att den håller samma höga kvalitet i förhållande till ett annat. Artikel I och II utgör exempel på validering i förhållandet till olika syften. I båda fallen handlar det om nationella prov i matematik, det vill säga olika syften med samma typ av prov.

Ett sätt att sammanfatta de viktigaste övergripande syftena med bedömningar återfinns i Nyström & Palm (2001b, sid. 41-42):

> *Insamling av information.* Detta är det primära och dominerande syftet med bedömningssituationer. Informationen som samlas in ligger till grund för olika beslut och omdömen, t ex betygsättning, elevens behov av repetition eller utvärdering av hur undervisningen fungerat.
>
> *Konkretisering av mål och kriterier*. Genom bedömningssituationerna utgår signaler om vad som är viktigt, och eleverna anpassar sitt lärande därefter. Om bedömningssituationerna verkligen speglar det som läraren anser vara viktigt kommer denna funktion att verka i positiv riktning, i annat fall kommer bedömningen att inverka kontraproduktivt på lärandet i den efterföljande undervisningen.
>
> *Lärande.* I de allra flesta bedömningssituationer finns ett mer eller mindre uttalat syfte att eleverna ska lära sig något av aktiviteten. Om bedömningssituationerna är utformade på sätt som är fördelaktiga för denna funktion blir den tid som läggs ner på bedömning mer effektiv.

När det gäller storskaliga bedömningar är syftet vanligtvis mer uttalat än vid bedömningar i klassrummet, och de svenska nationella proven är ett aktuellt svenskt exempel (Skolverket, 2002, 2003). Det övergripande syftet med kursproven är att de ska implementera måldokumenten och bidra till att öka likvärdigheten i betygssättningen över landet.

> Det nationella provsystemet syftar till att ge lärare stöd vid diagnostisering och betygssättning av elevernas kunskaper för att därigenom skapa förutsättningar för att bedömningsgrunderna ska bli så enhetliga som möjligt över landet.
> (http://www.skolverket.se, 2003-05-20).

Vidare har den rapportering av resultat från nationella prov som Skolverket gör som ett syfte att presentera en nationell resultatbild som kommuner,

skolledare och lärare ska kunna använda vid utvärderingar av den egna verksamheten. Avsikten är också att ge ett underlag för jämförelser mellan lärares egna resultat och resultaten för ett urval av gymnasieskolor och komvuxenheter. Skolverket samlar in provresultat från ett antal skolor och detta anges att det kommer att användas i olika utvärderingar av svensk skola och för forskningsändamål. Materialet är även ett viktigt underlag för arbetet med utveckling av proven.

Ovanstående speglar intentionerna med de nationella proven i Sverige, men vilka funktioner och roller har de nationella proven i realiteten? Vi kan konstatera att det inte tycks finnas mycket av systematiskt vetande i frågan. Det som finns är undersökningar av hur lärare och andra som arbetar i skolan uppfattar proven. I Skolverkets utredning av det nationella provsystemet (Skolverket, 2003) konstateras till exempel att lärare uttrycker uppfattningar om att proven riskerar att bli alltför styrande för undervisningen och hämmande för pedagogiskt nytänkande, men det finns också uppfattningar om att proven tvärtom har en viktig funktion i det senare avseendet. En majoritet av lärare tycker dock inte att proven styr undervisningen i för hög grad.

Internationellt finns det inte heller mycket empiriska belägg för hur storskaliga bedömningar, till exempel nationella prov, påverkar undervisning. Det finns en mängd antaganden och anekdotiska belägg (Wiggins, 1993) men bara ett fåtal mer systematiska studier (t.ex. Barnes, Clarke, & Stephens, 2000). Mehrens (2002) konstaterar att bristen på sådana studier är påtaglig i USA. Han framför dock hypotesen att storskaliga bedömningar påverkar i högre grad om de är avgörande för till exempel betyg eller intag till högre studier (så kallad ”high-stake assessment”). En studie från Nya Zeeland visar att storskaliga bedömningar kan ha positiva effekter om de åtföljs av ett program för lärarfortbildning (Gilmore, 2002). Broadfoot (2002) konstaterar att bedömning generellt antas påverka undervisning i hög grad, men att det är

> …surprising how little systematic attention that appears to be being given to evaluating these effects or indeed to questioning the capacity of the tools being employed to deliver what is being sought. (sid. 286)

Hon frågar sig om all möda som läggs ned på bedömning verkligen bidrar till kvalitativt bättre undervisningsresultat.

Syften med bedömningar kan vara uttalade, men de kan också vara dolda. Dessutom är det inte alltid så att avsedda syften överensstämmer med den

användning och påverkan som bedömningar får. Under alla omständigheter är de nödvändigt att reflektera över såväl uttalade som dolda syften för att en validering ska vara möjlig. Ett annat område som utgör en nödvändig utgångspunkt för validering av bedömning i skolan är läroplanens mål.

Läroplanens mål

Läroplanen är på många sätt relevant för bedömningsfrågorna, det vill säga frågorna om varför bedömning behövs, vad som ska bedömas, vem som ska bedöma och hur bedömning kan ske. Följaktligen är läroplanen också viktigt vid en värdering av kvaliteten hos bedömningar. Det är uppenbart att läroplanens mål bör ligga till grund för bedömningar i skolan. Avsikten med detta avsnitt är att visa på några relevanta aspekter när det gäller synen på läroplan i allmänhet och mål för utbildningen i synnerhet.

Vad är en läroplan?

När jag påstår att en läroplan är nödvändig för att värdera kvaliteten i bedömningar så sker det utifrån en syn på läroplan som nära ansluter till Lundgren (1979) som skriver att en läroplan ”utgör samhällets krav på fostran och utbildning” (sid. 231). Observera att denna läroplan innehåller såväl övergripande mål, till exempel värdegrund, som mer specifika mål, till exempel kursplaner.

Olika aspekter av läroplaner

Med läroplan avses ofta de styrdokument som reglerar skolans pedagogiska verksamhet. Ett besläktat begrepp är ”curriculum” som dock ofta representerar inte bara mål för skolan, utan även metoder för undervisning och ibland till och med genomförandet av skolarbetet. Såväl läroplan som curriculum är dock något mer än det som finns skrivet i nationella styrdokument eller lokala arbetsplaner. Det finns andra mer svårfångade aspekter av läroplansbegreppet. Graybill (1998) beskriver med hänvisning till Sinclair & Ghory (1987) tre aspekter av en ”curriculum”: den uttryckta (expressed), den antydda (implied) och den framväxande (emergent). Den uttryckta läroplanen representerar de texter som innehåller mål och aktiviteter vilka fastställs på nationell nivå och implementeras av lärare. Den antydda läroplanen är de hemliga budskap som åtföljer den uttryckta läroplanen, det vill säga den icke uttalade eller dolda läroplanen. Den framväxande läroplanen är de pågående förändringar och justeringar som görs i den uttryckta och antydda läroplanen för att koppla det unika hos varje elev med läroplanen. I en annan uppdelning talas det om den avsedda, implementerade respektive erfarna läroplanen (Mullis et al., 2003). Sett i bedömningsperspektiv har alla dessa aspekter av läroplanen betydelse. Till exempel är en värdering av rättvisan i en bedöm-

ning mer beroende på den framväxande läroplanen än på den uttryckta. Det intressanta är hur läroplanen kommer till uttryck när den samverkar med till exempel lärares föreställningar och organisatoriska ramar.

Den svenska läroplanen

Sivesind, Bachman, & Afsar (2003) analyserar de nordiska läroplanstexterna och använder sig av fyra olika läroplanstyper för att kunna kategorisera och jämföra såväl gällande läroplaner som den utveckling av läroplaner som kan anas. De fyra läroplanstyper som de definierar är (1) läroplanen som ett formellt reglerande dokument, (2) läroplanen som ett auktoriserande och innehållsbeskrivande dokument, (3) läroplanen som ett politiskt normerande dokument, och (4) läroplanen som ett standardiserande dokument för central utvärdering av utbildningen. Sivesind et al. (2003) konstaterar i sin analys av den svenska läroplanen att den närmast kan karakteriseras som läroplanstyp 3. Det innebär att både läroplanen (inklusive kursplaner) närmast kan karakteriseras som politiskt normerande dokument genom att överväganden om mål och riktlinjer sätts in i en samhällelig kontext, där sociala aspekter har ”en långt mer framträdande plats än pedagogiska och didaktiska aspekter” (sid. 22). I ett bidrag till Skolverkets bedömning av dagens svenska system med nationella prov, med avseende på kvalitet och kostnadseffektivitet (Skolverket, 2003) hänvisas också till den norska kategoriseringen av läroplaner (Sivesind et al., 2003), och Skolverkets arbetsgrupp konstaterar att en tillämpning av denna kategorisering på svenska förhållanden påvisar påtagliga spänningar i systemet.

> För att stärka kontrollen och insynen har inspektion införts, vilket närmast för tankarna till läroplanstyp 1 eller 2, beroende på vad som ska inspekteras. De nationella provens betydelse har samtidigt förskjutits från att vara ett betygsstödjande instrument med fokus på lärarens yrkesutövning, mot att bli ett utvärderingsinstrument för olika administrativa nivåer. Från att vara ”kunskapsbedömning *för* lärande” mot att bli ”kunskapsbedömning *av* lärande”. Från att främst ha en formativ roll mot att ha en summativ. Från att vara kunskapsbedömning enligt typ 3 mot att bli kunskapsbedömning enligt läroplanstyp 4. (Skolverket, 2003, sid. 122)

En sådan perspektivförskjutning har relevans för valideringen av de nationella proven.

En annan läroplansaspekt handlar om relationen mellan generella mål och mål i olika skolämnen. I en kommentar till läroplan, kursplaner och betygskriterier skriver Skolverket:

> Man läser inte ämnen så mycket för att lära sig särskilda fakta och begrepp utan för att lära sig uppfatta saker och använda begrepp på särskilda sätt. Genom de olika ämnena erövrar man de särskilda sätt att erfara och förhålla sig till världen som utvecklats inom de kunskapstraditioner som enskilda ämnen eller ämnesgrupper representerar. (Skolverket, 1996, sid. 2)

Vidare skriver man att det enskilda innehållet i ett ämne inte är oviktigt, utan att det är en förutsättning för att utveckla olika kunskapskvaliteter.

> Man kan inte utveckla kvaliteterna i sig, lära sig tänka i största allmänhet eller lösa problem i största allmänhet. Det är alltid något man tänker om och något specifikt problem man löser. På så vis är det konkreta innehållet en förutsättning för utvecklingen av de mer generella kvaliteterna. (ibid., sid. 2).

Jag menar att det är viktigt att hitta en balans mellan mål i ett ämne och de generella mål som studier i ämnet kan bidra till. Hur denna balans ser ut får betydelse för vad som är viktigt att bedöma och också viktigt för hur bedömningssituationer utformas.

De svenska nationella styrdokumenten för skolan anger mål på olika nivåer, men avsikten är att enskilda mål ska förstås utifrån andra mål. ”De olika mål- och styrdokumenten utgör en helhet och de skall läsas tillsammans” (Skolverket, 1996, sid. 6). En väsentlig skillnad mellan olika mål i styrdokumenten är dock att de riktar sig till olika aktörer i skolan. I läroplanen (Utbildningsdepartementet, 1994) anges mål som riktar sig till skolan, till exempel ”Skolan skall sträva efter att varje elev …” och ”Skolan ansvarar för att varje elev efter genomgången grundskola …”. Även programmålen för gymnasieskolans olika program riktar sig till skolan: ”Skolan skall ansvara för att eleverna vid fullföljd utbildning … (se till exempel Skolverket, 2000). I kursplanerna för varje ämne anges två olika typer av mål. När det gäller mål att sträva mot så riktar de sig också till skolan (”Skolan skall i sin undervisning sträva efter att eleven …). Det är endast de så kallade målen att uppnå som direkt riktar sig till eleven. I gymnasieskolans kursplaner, till exempel, så anges mål som eleverna skall ha uppnått efter avslutad kurs i olika satser som inleds med ”Eleven skall …”.

Vad betyder detta för bedömningsfrågorna? Som styrdokument tycks det endast vara mål att uppnå som eleverna i en bedömningssituation så att säga

kan ställas till svars inför. Om elever inte kommit särskilt långt i en riktning som pekas ut av mål att sträva mot så kan möjligen skolan anklagas för att inte ha följt styrdokumenten, men den enskilda eleven bör inte straffas för det. Detta innebär att bedömning av enskilda elevers måluppfyllelse enligt styrdokumenten endast kan grundas på mål att uppnå. Samtidigt är intentionen med styrdokumenten att de ska läsas tillsammans. Det måste till exempel innebära att tolkningen av mål att uppnå inte ska harmoniera med andra mål, riktlinjer och idéer som styrdokumenten står för. En värdering av hur värdegrund, mål som riktar sig till skolan och mål som riktar sig till enskilda elever hänger ihop och samverkar är viktig för att avgöra hur och på vad som bedömningen ska inrikta sig.

Kunskapssyn

För validering av bedömningar i skolan är även kunskapssynen central. Läroplanen säger en del om kunskapssyn, men än viktigare är föreställningarna hos dem som utformar och använder bedömningar i skolan som kommer till uttryck i de bedömningar som görs.

Jag vill i det följande kortfattat stanna vid tre beröringspunkter mellan kunskapssyn och bedömningar. För det första har våra uppfattningar om lärande och vilken kunskap som eftersträvas betydelse för bedömningar i skolan. För det andra betyder uppfattningen om hur olika nivåer av kunskap konstrueras mycket för hur bedömningar utformas och genomförs. För det tredje är föreställningar om hur vi kan bilda kunskap om andras kunskap och kompetens viktiga.

Vad är kunskap och lärande?

I artikel IV redovisas hur en av de intervjuade lärarna liknar undervisning vid att bygga en vägg med tegelstenar. Metaforen tyder på en uppfattning av kunskap som bestående av isolerade delar som kan läras var för sig. Man kan fråga sig hur en valid bedömning ser ut som bygger på en sådan kunskapssyn?

Rådande uppfattningar om kunskap och lärande har förändrats över tid. Kunskapssynens utveckling under 1900-talet, och dess relation till bedömningsfrågorna har analyserats av bland andra Caroline Gipps och Lorrie Shepard (Gipps, 1994; Shepard, 2000). För hundra år sedan dominerades scenen av teorier om lärandet som byggde på behaviorism och associationism, där lärandet betraktades som en ackumulering av stimulus-respons associationer. (Thorndike, Skinner, med flera). Nu, hundra år senare, finns en variation i hur olika aktörer ser på saken, men den officiella synen på kunskap och lärande är i alla fall radikalt annorlunda jämfört med den ovan beskrivna synen som domineras av behaviorism och associationism.

Enligt Shepard (2000) uppfattas lärandet nu som en aktiv process som kännetecknas av mentala konstruktioner och begripliggörande, i kontrast till tidigare mekanistiska teorier om lärande. Dessutom framhåller kognitiv teori att existerande kunskapsstrukturer och föreställningar har stor betydelse för att möjliggöra eller hindra nytt lärande, att kvalificerat tänkande innefattar självvärdering och en medvetenhet om när och hur färdigheterna ska användas, samt att expertis utvecklas inom ett fält som ett principfast och koherent

sätt att tänka och representera problem och inte bara som en ackumulering av information. En annan viktig idé inom detta ”paradigm” är att utveckling och lärande primärt är sociala processer.

Linde (2003) gör en intressant jämförelse mellan kunskapssynen i olika ämnen så som den framträder i grundskolans och gymnasieskolans kursplaner. Han menar att de ämnen som han jämför (engelska, matematik och samhällskunskap och religionskunskap) i många avseenden är olika i sina epistemologiska grundvalar. Däremot är skillnaderna mellan grundskola och gymnasium små och enligt Lindes analys går skiljelinjen främst mellan ämnen, inte mellan stadier och åldersgrupper i skolsystemet. I engelska värderas framförallt en praktiskt orienterad kunskap i form av kommunikation, det vill säga kunskap som är värdefull för sin praktiska nyttas skull. När det gäller matematik så talar kursplanerna om såväl den praktiska nyttan som nödvändigheten av vetenskapsteoretiska reflektioner över matematiken som en av människan konstruerad artefakt. Här handlar det alltså även om kunskap som har ett värde för att ”utveckla en högre insikt, en kritisk, filosofisk betraktelse av vår kunskap och dess ursprung och grund” (sid. 70). Linde (2003) beskriver en kunskapssyn i matematik som karakteriseras som konstruktivistisk, och kursplanen tycks därmed starkt influerad av ett sätt att se på matematikämnet och lärande i matematik som på senare år kommit att dominera diskursen bland forskare och läroplansteoriker (se t.ex. Romberg, 1992).

Denna skillnad mellan kunskapssynen i matematik och engelska lyfter Skolverket fram som en tänkbar förklaring till de skillnader mellan resultatbilderna på nationella prov som kan konstateras (Skolverket, 2003). Här gör alltså Skolverkets utredare en koppling mellan kunskapssyn och bedömning.

Om vi utgår från att olika kunskapssyner framträder i kursplanerna för olika ämnen, hur ska det värderas? Ska sådana skillnader betraktas som misslyckanden i processen att skriva kursplaner? Eller kan det vara så att ämnenas olika karaktär också innebär att de präglas av olikheter i kunskapssynen? Det är angeläget att explicitgöra och kritiskt granska för givet tagna ”sanningar” som att alla ämnen är lika i detta avseende, eller att de nivåer av kunskap (kvalitativt olika sätt att erfara världen) som återspeglas i betygskriterier för olika ämnen skulle kunna vara jämförbara. Det är inte minst viktigt för att kunna reflektera över kvaliteten i de bedömningar som används i skolan.

Det bör också understrykas att Lindes analys sker utifrån kursplanetexten. Andra analyser måste till för att studera om detta också är den bild av väsentliga kunskaper i ämnet som förmedlas i undervisning eller i bedömningssituationer.

Nivåer av kunskap

Bedömning i skolan kan sägas gälla kunskapskvaliteter. Ofta betecknas dessa i termer av bättre och sämre. Det gäller till exempel betygssystemet där kriterier för olika betyg ska spegla olika kvaliteter i elevernas kunnande (Skolverket, 1996), men det faktum att ett Väl godkänt betyg ges ett större värde än ett Godkänt betyg innebär att kriterierna utgör en hierarki.

Ett antal försök att konstruera hierarkier av kunskaper har presenterats och använts. Några exempel är Blooms taxonomi (Bloom, 1956), SOLO-taxonomin (Biggs & Collis, 1982) samt en relativt ny revision av Bloom (Andersson et al., 2001). Det svenska skolsystemets nationella betygskriterier är ett annat exempel. En beskrivning och diskussion av flera hierarkier, inklusive de nationella betygskriterierna, med speciellt fokus på prestationer i matematik återfinns i Nyström (1998). Det är viktigt att betona att varje sådan konstruktion är historiskt och kontextuellt betingad. Det finns inga absoluta kriterier för olika nivåer av kunskap. I Nyström (1998) visas också hur uppgiftens karaktär samverkar med de kriterier för nivåer av kunskap som används vid bedömningen. En uppgift kan inbjuda till prestationer som betraktas ha mycket hög kvalitet utifrån ett sätt att se på nivåer av kunskap, medan uppgiften inte inbjuder till en typ av prestationer som betraktas ha mycket hög kvalitet utifrån ett annat sätt att se. Uppgiftens karaktär spelar alltså roll för vilka prestationer som kan förväntas. Samtidigt är det inte i allmänhet så att prestationen bestäms av uppgiften. I alla bedömningssammanhang är det prestationen som är av en viss kvalitet. Ett undantag gäller de mest strukturerade uppgifterna, det vill säga uppgifter med fasta svarsalternativ. I det fallet styrs prestationens komplexitet i mycket hög grad av hur svarsalternativen valts och formulerats. Med andra ord kan uppgifter i allmänhet inte kopplas till en högre eller lägre nivå av kunskap, hur än dessa nivåer formuleras, även om det finns uppgifter som i olika grad stimulerar elever att visa prestationer på olika nivåer.

Hur kan kvaliteten hos individers och gruppers kunskaper studeras?

Frågan om hur kunskap hos andra kan studeras är naturligtvis en grundläggande fråga eftersom det är vad bedömning av kunskap och kompetens ytterst handlar om. Vilka olika ställningstaganden kan man tänka sig i frågan om hur vi kan få kunskap om andras kunskap och kompetens? Vi kan utgå från en självdefinierande bedömning, där det är det som någon uttrycker (muntligt, i skrift eller i handling) vid en viss tidpunkt som räknas, oavsett om det representerar något bakomliggande mönster eller inte. Eftersom det som bedöms definieras genom bedömning så blir frågan om hur vi kan veta något om andras vetande oproblematisk. Problemet försvinner också om man föreställer sig att det som människor presterar alltid visar deras kompetens. Detta är kännetecknande för en positivistisk kunskapssyn, som enligt Guba & Lincoln (1994), kännetecknas av dualism och objektivism. Det studerade ”objektet” och den som studerar antas vara oberoende enheter, och den undersökande antas vara kapabel att studera objektet utan att påverka det eller låta sig påverkas. När påverkan i någondera riktning upptäcks eller misstänks så används någon strategi för att minska eller avlägsna denna påverkan mellan subjekt och objekt. En radikalt annorlunda utgångspunkt är att var och en konstruerar sin kunskap, vilket ytterst gör det omöjligt, eller åtminstone svårt, för någon annan att göra sig en bild av individens kunskap (se till exempel Björkqvist, 1993).

Med utgångspunkt i antagandet om att det går att få reda på en del om elevers kunskap och kompetens väcks förstås frågan om hur detta kan ske. Vi ska här inte gå in närmare på olika former för bedömning, det har gjorts på många andra håll, se till exempel Nyström & Palm (2001a; 2001b). Ofta präglas diskussioner om hur bedömning kan ske av en kategorisering i ”traditionella” och ”innovativa” bedömningsformer. Användningen av sådana värdeladdade beskrivningar riskerar att dölja viktiga kvalitetsaspekter när det gäller bedömningar, och det finns goda skäl att undvika uppdelningar och etiketteringar av det slaget.

> We need to move the debate away from false dichotomies: criterion-referenced assessment versus norm-referenced assessment, standardized tests versus performance assessment; they are as unhelpful as the quantitative *versus* qualitative research method argument. What we need is to understand the value of each approach and to follow the fitness for purpose principle; some approaches then, of course, may be seen to have little value. (Gipps, 1994, sid. 163)

Varje ansats att lära sig om någon annans kunskap och kompetens måste värderas utifrån rimliga kriterier, och jag menar att den modell som beskrivs i denna avhandling kan vara en utgångspunkt för en sådan värdering.

Sammanfattning

Sammanfattningsvis menar jag att Messicks ramverk för validitet (se Tabell 1) utgör en kraftfull modell för att både reflektera över kvalitet i bedömningar i skolan och att utveckla bra bedömningssituationer. Messicks strukturering av ett enhetlig och omfattande validitetsbegrepp ger en teoretisk utgångspunkt för att värdera de olika delarna av bedömningsprocessen (se Figur 1). Den teoretiska modellen behöver dock fyllas med relevanta och konkreta validitetsfrågor. Jag har pekat på ett antal exempel på kriterier för värdering av bedömningar, och jag menar att dessa kan koppla ihop teori och praktik.

Jag har argumenterat för att syftet med en bedömning är en oundgänglig utgångspunkt vid validering, och att om en viss bedömning har flera syften så måste validering ske gentemot varje syfte för sig. Vidare implicerar Messicks validitetsbegrepp att såväl läroplan som kunskapssyn måste explicitgöras för att valideringen ska bli meningsfull. Med läroplan avses inte bara den avsedda eller uttryckta läroplanen, utan också den implementerade och erfarna. Med andra ord är bedömningens syfte, läroplanens mål och kunskapssynen nödvändiga utgångspunkter för validering av bedömningar i skolan.

Avhandlingens fyra artiklar har bidragit på olika sätt till det övergripande syftet att utveckla en utgångspunkt för värdering av kvalitet i samband med bedömningar i skolan. Till exempel demonstreras i artikel I hur Messicks validitetsbegrepp kan användas för att påvisa hur olika aspekter av validitet kan motverka varandra. Samtidigt har utläggningen om validitet och dess förutsättningar i syfte, läroplan och kunskapssyn avsett att belysa en av de viktigaste utgångspunkterna för de studier som sammanfattas i följande avsnitt.

Sammanfattning av artiklarna

I Reliability of Educational Assessments – The case of classification accuracy.

Artikeln handlar om reliabilitetsproblemet, det vill säga risken för instabilitet och slumpmässighet i samband med bedömningar. Hur stor är till exempel risken att olika bedömare ger olika omdömen för samma prestation? Eller hur stor är risken att en elevs prestation är påverkad av slumpmässigheter som att personliga problem har påverkat möjligheterna att visa vad man kan?

Problemet finns vid alla bedömningar, och betraktas i denna studie som en aspekt av validiteten, i enlighet med Messicks validitetsbegrepp. Ofta sätts likhetstecken mellan reliabilitetsproblemet och ett eller annat mått på problemets omfattning (”reliabilitetskoefficient”). Det är angeläget att dessa skiljs åt. Det är också angeläget att hitta begrepp och metoder som på ett meningsfullt sätt kan göra det möjligt att skatta storleken på reliabilitetsproblemet. I artikeln föreslås att ”classification accuracy” (CA) är ett sådant begrepp.

Begreppet bygger på tankegången att en individ kan vara kompetent på ett område, men att eleven inte nödvändigtvis lyckas visa det i en bedömningssituation. På motsvarande sätt kan vi föreställa oss en individ som inte är kompetent, men som av någon tillfällighet tycks motsvara kraven vid en bedömning. CA representerar sannolikheten att de som är kompetenta också visar kompetenta vid bedömningen samt att de som inte är kompetenta inte heller tycks kompetenta vid bedömningen.

I artikeln redovisas skattningar av CA för ett svenskt nationellt prov som ett exempel på hur reliabilitetsproblemet kan vara relevant för valideringsprocessen. Två metoder används för att skatta denna ”accuracy”; dels en metod hämtad från Livingston & Lewis (1995) och dels en som jag själv utvecklat utifrån klassisk testteori.

Det konkreta problem som studeras är hur reliabiliteten påverkas av olika modeller för kravgränser i ett nationellt kursprov i matematik. Om provet delas in i olika innehållsliga domäner och prestationen i varje domän poängsätts kan prestationskrav ställas i form av kravgränser i varje domän. Detta styrker bland annat möjligheterna att koppla provresultatet till utbildningens mål, och ger på så sätt ett positivt bidrag till provets validitet. Samtidigt innebär dock en sådan kravgränsmodell att reliabiliteten riskerar att minska

eftersom varje delprestation (på varje domän) bedöms utifrån ett färre antal uppgifter. Genom att sätta kravgränser på provet som helhet ökar antalet uppgifter som kategoriseringen i Godkänd respektive Icke godkänd, och därmed kan reliabilitetsproblemen minska. Samtidigt innebär sådana kravgränser en försämring av andra validitetsaspekter.

CA skattas vid tillämpning av de två modellerna för kravgränser på resultat från 1201 elever på ett nationellt kursprov i matematik. De skattningar av reliabilitetseffekterna som redovisas i artikeln antyder att reliabiliteten påverkas i relativt liten utsträckning av de olika sätten att ange kravgränser. Vad som betecknas som små reliabilitetsförändringar är dock en fråga om värdering.

II National tests as a means for evaluating effects of streaming in upper secondary school mathematics.

I denna studie kopplas frågor om bedömningars funktion och validitet ihop med frågor om hur gruppering av elever påverkar deras prestationer.

Den svenska gymnasieskolan består av 17 olika program som har olika inriktningar. På grund av programmens karaktär och på grund av elevernas val så innebär denna programstruktur en gruppering av elever utifrån tidigare prestationer i skolan. Gymnasieskolans programstruktur karakteriseras av skillnader mellan program, men också av likheter i form av kurser som alla elever måste gå, de så kallade kärnämneskurserna. Det innebär att alla elever ska bedömas utifrån samma mål och samma betygskriterier i dessa kurser. En sådan kurs är Matematik A, som dessutom avslutas med ett obligatoriskt nationell kursprov som är gemensamt för alla program.

Forskning har visat att om elever med till synes samma prestationsnivå placeras i olika nivågrupper i matematik kommer att visa olika prestationsnivåer efter en tid. De elever som placeras bland elever med relativt hög prestationsnivå kommer att prestera bättre. En utgångspunkt för denna studie var att undersöka om detta fenomen också kan observeras i den svenska gymnasieskolan, och om resultat från nationella prov kunde användas i detta syfte.

Studiens syften är alltså att undersöka möjligheterna att använda resultat från nationella prov för att studera elevernas kunskapsutveckling i matematik och att undersöka effekterna av elevernas programval på deras prestationsnivå i matematik. En databas upprättades där 403 elevers resultat på det nationella provet i Matematik A kunde matchas mot samma elevers resultat på

ämnesprovet i skolår 9. Med hjälp av bland annat linjär regression studerades sambandet mellan provprestationerna på de två proven för elever på olika program.

Resultatet visar att prestationen är avsevärt bättre för elever som grupperats tillsammans med högpresterande elever. Denna effekt tycks vara störst för lågpresterande elever. Slutsatsen att effekten inte är lika stor för högpresterande elever är osäker på grund av så kallade takeffekter. Provet är helt enkelt inte konstruerat för att skilja de mest högpresterande eleverna åt.

De skillnader som observeras kan utifrån litteraturen förstås som att elever som grupperas med högpresterande elever får en positiv utveckling på grund av stimulerande klassrumsmiljö, högre förväntningar från läraren och en utökad kurs.

Sammanfattningsvis visar studien hur nationella prov kan användas för utvärderingsändamål, men att denna användning också har avsevärda begränsningar och problem. Vidare ger studien intressanta resultat när det gäller skillnader mellan program.

III Students cues to perceived competence in mathematics.

Denna studie handlar om elevers upplevda kompetens. Den upplevda kompetensen, och andra närliggande, har i många studier visat sig påverka elevers val av utbildning, hur mycket man anstränger sig i olika situationer och till och med hur väl man presterar. Det finns en stor mängd forskning som visat på hur viktig den upplevda kompetensen är.

Studien fokuserar frågan om vilka källor som elever använder sig av för att forma sin upplevelse av kompetens i matematik och sin bild av andras kompetens. Dessutom undersöks om den skillnad mellan kvinnor och män avseende upplevd kompetens som denna och många andra studier påvisar kan kopplas till att kvinnor och män använder olika källor för att bilda sin upplevda kompetens.

Studien bygger på en enkätundersökning som gjordes i samband med ett nationellt kursprov i matematik våren 2002. Enkäten, som bestod dels av frågor om provet och dels av 23 frågor som handlade om upplevd kompetens. Resultaten bygger på enkätsvar från 521 elever.

Eleverna fick först svara på frågan Hur bra är du i matematik? Eleven hade fem svarsalternativ att välja på, från Mycket bra till mycket dålig. Denna fråga, samt frågor om vad eleverna trodde att de skulle få för betyg på provet och på kursen, användes som mått på elevernas upplevda kompetens.

En grupp frågor handlade om vilka källor som eleverna använder för att skapa en bild av sin egen kompetens och hur viktiga dessa källor är. Enligt enkätsvaren tycker eleverna som grupp att en känsla av motivation och förståelse ger det viktigaste till den upplevda kompetensen. På andra plats kommer mer objektiva källor som provresultat och lärarens omdöme om hur bra det går. Minst viktiga för upplevelsen av kompetens är faktorer som har att göra med arbetet i klassrummet, till exempel hur många uppgifter som eleverna hinner lösa.

Svaren visar inte på några könsskillnader, dvs. kvinnor och män värderar olika källor på likartat sätt. Den lägre upplevelsen av kompetens bland kvinnor som konstaterats i flera studier kan alltså inte på grundval av denna studie sägas ha med användning av olika källor att göra. Det är snarare så att samma källor används, men att de används på olika sätt.

IV Lika barn leka bäst? En gymnasielärardiskurs om nivågruppering i matematik.

Problematiken kring nivågruppering i den obligatoriska skolan har ägnats stort intresse från forskares sida, och en stor mängd forskningsrapporter av olika slag har presenterats genom åren. När det gäller nivågruppering av elever från sextonårsåldern och uppåt finns det däremot väldigt lite forskning publicerad. Syftet med denna studie är att beskriva och tolka hur matematiklärare vid en gymnasieskola uttrycker sina erfarenheter av och föreställningar om inre differentiering (nivågruppering) i en skolform som präglas av en stark yttre differentiering avseende nivå, kön, intresse och ambition. Vilka motiv, möjligheter, nackdelar och svårigheter ser lärare i nivågruppering?

Studien begränsades till en gymnasieskola, och sex matematiklärare intervjuades. Den handlar om lärares föreställningar och tar sin utgångspunkt i en form av diskursanalys (se Potter & Wetherell, 1987). Avsikten var att spegla variationen i hur matematiklärarna vid skolan pratar om nivågruppering.

Resultatredovisningen struktureras kring fyra teman, som har sitt ursprung i de teman som låg till grund för intervjuerna. Det först temat handlar om elevers olikheter och vad dessa olikheter har med nivågruppering att göra. Det andra temat behandlar lärarnas syn på nivågrupperingens fördelar, nackdelar, möjligheter och brister. Det tredje temat berör hur undervisning kan och bör skilja sig åt mellan elevgrupper på olika ”nivåer”. Det fjärde och sista temat handlar om vad lärarna tror att andra tycker om nivågruppering.

Sammanfattningsvis visar lärarnas uttalanden upp ett brett spektrum av nackdelar och problem med nivågruppering. Samtidigt visar lärarrösterna ganska stor samstämmighet när det gäller till exempel en grundläggande positiv inställning till den form av nivågruppering som tillämpats vid skolan, att problem med nivågruppering i första hand gäller så kallade svagare eller långsammare grupper, att blandningen av omotiverade elever och elever som tycker att matematik är svårt är problematisk, och att skillnaderna mellan hur undervisning bedrivs i grupper på olika nivåer är ganska små. I intervjuerna kommer lärarna också in på bedömningsfrågorna. Till exempel uttrycker en lärare att eleverna upplever sig bli bedömda på olika grunder i olika grupper.

Min tolkning är att lärarna försöker hitta en väg genom en komplex verklighet med motstridiga mål och yttre begränsningar. De upplever att vissa elever behöver och vill ha mer matematik än vad som normalt erbjuds på det gymnasieprogram de valt, och att många andra elever inte når målen för kursen om de inte får mer tid och hjälp. Lärarna ser nödvändigheten att göra något åt situationen och nivågruppering ligger nära till hands, i synnerhet mot bakgrund av de ramar i form av lokaler och gruppstorlek som lärarna beskriver och mot bakgrund av lärarnas föreställningar om matematik och matematikundervisning.

Ett intressant resultat är de relativt små skillnader som lärarna uttrycker i sina beskrivningar av undervisning i klassrum med hög- resp. lågpresterande elever. Detta är inte en klassrumsstudie, det är en studie av lärares föreställningar och diskurs, men detta resultat kan tolkas som en indikation om hur praktiken ser ut. Under alla omständigheter antyder detta resultat att lärarna inte har en föreställning om en sådan variation i matematikundervisning och klassrumsaktiviteter som skulle kunna ta vara på de möjligheter som eventuellt kan skapas genom en nivågruppering.

Om vetenskapliga ansatser och metodiska överväganden

Olika ansatser och metoder finns representerade i det empiriska arbete som redovisats i denna avhandling. Jag ska i det följande försöka beskriva utgångspunkterna för denna variation.

Olika ansatser

Utgångspunkten i artikel I-III kan närmast betecknas som postpositivistisk. Guba & Lincoln (1994) beskriver postpositivismens ontologi som en kritisk realism. En verklighet antas existera, men denna verklighet kan endast begripas ofullständigt på grund av grundläggande brister i de mänskliga intellektuella mekanismerna och de studerade fenomenens fundamentalt motspänstiga natur. Påståenden om verkligheten måste utsättas för omfattande kritiska prövningar för verkligheten ska kunna beskrivas och förstås så nära som möjligt, men denna förståelse kan aldrig bli fullständig eller perfekt. Epistemologin hos postpositivismen betecknar Guba & Lincoln som ”modifierat dualistisk/objektivistisk”. Dualismen anses inte möjlig att upprätthålla, men objektivitet kvarstår som ett styrande ideal. Replikerbara resultat är förmodligen sanna, men alltid öppna för falsifikation. Metodologin syftar till att motverka en del av kritiken mot positivismen genom att undersökningar görs i mer naturliga omständigheter, med insamling av mer information om situationen och ett återinförande av upptäckandet som en del av undersökningen. I samhällsvetenskap i synnerhet används ”objektens” röster för att hjälpa till att bestämma meningar och syften som människor kopplar till sitt agerande. Alla dessa mål åstadkoms i huvudsak genom en ökad användning av ”kvalitativa” tekniker.

Ansatsen i artikel IV bygger på en form av diskursanalys, vilket innebär att utgångspunkterna är delvis annorlunda än i den postpositivistiska ansats som beskrivits ovan. Diskursanalys, i den socialpsykologiska tappning som presenterats av Potter & Wetherell (1987), använder ett relativt snävt definierat diskursbegrepp. Med diskurs avses varje form av språkligt uttryck, det vill säga tal och skrift. Diskursanalysens ontologi är mer idealistisk än postpositivismens, här görs inga antaganden om att någon verklighet måste finnas. Språkets konstruerade och konstruerande karaktär är viktig, liksom dess variation.

> We are not trying to recover events, beliefs and cognitive processes from participants' discourse, or treat language as an indicator or signpost to some other state of affairs but looking at the analytically prior question of how discourse or accounts of these things are manufactured (Potter & Wetherell, 1987, sid. 35).

Utgångspunkten är att människor lever och erfar i en diskurs, det vill säga att diskursen sätter ramar för vad som kan erfaras eller hur erfarenheter uppfattas, och påverkar därigenom vad som kan sägas och göras. (Potter & Wetherell, 1987)

Epistemologin har inte samma dualistiska och objektivistiska prägel som i postpositivismen, utan gränsen mellan forskningens subjekt och objekt är mycket svagare.

Sammanfattningsvis använder jag mig av två olika ansatser som skiljer sig åt när det gäller såväl ontologi som epistemologi. Samtidigt ska inte dessa skillnader överdrivas. Till exempel menar jag att diskursanalysens ontologi i hög grad är förenlig med postpositivismens.

Olika metoder

Metoderna som används i artikel I-III är dels enkäter och dels prov där resultaten sammanfattas med kvantitativa mått. Artikel IV bygger på intervjuer. Det är med andra metoder som ofta betecknas som kvantitativa respektive kvalitativa, beteckningar som jag helst undviker. Jag menar i enlighet med (Åsberg, 2001, sid. 270) att

> [f]rågan om vad för slags kunskap vi producerar genom vår forskning behöver diskuteras befriad från det kvantitativa-kvalitativa argumentets hämmande och missvisande retorik.

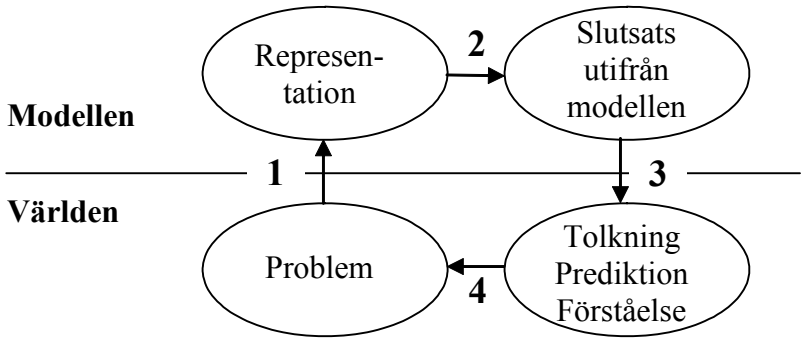
Det finns flera olika argument för att uppdelningen i kvalitativ och kvantitativ metod är missvisande. Ett viktigt skäl är att all forskning som bygger på ”kvantitativa” metoder i grund i grund och botten är kvalitativ. Detta kan exemplifieras med enkätundersökningar, som ofta betecknas som ”kvantitativ” forskning eftersom resultaten sammanfattas i siffror. Användning av enkäter kräver dock bland annat att forskaren väljer lämpliga frågor i förhållande till forskningens syfte, väljer lämpliga svarsalternativ eller väljer att efterfråga egenproducerade svar, samt formulerar frågor och svarsalternativ på lämpligt sätt Allt detta handlar i högsta grad kvalitativa överväganden.

Samtidigt som all ”kvantitativ” forskning i grund och botten är ”kvalitativ” så finns ofta kvantitativa inslag i så kallad kvalitativ forskning. Åsberg

(2001) konstaterar att det inte finns något sådant som kvalitativ eller kvantitativ metod. Enligt Åsberg hänvisar kvantitativ och kvalitativ till egenskaper hos de fenomen vi söker kunskap om. Det skulle betyda att data kan vara kvalitativa (i form av ord) eller kvantitativa (i form av siffror). Men kvalitativa data ger ibland upphov till kvantitativa resultat. Även den som säger sig vara ”kvalitativ” gör då och då jämförelser av kvantitativ natur. Till exempel är det inte ovanligt att hitta kvantitativa jämförelser som ”mer än” eller ”ofta”, i forskningsrapporter som säger sig vara ”kvalitativa”.

Denna avhandlings titel anspelar på mätning som en utgångspunkt för att studera fenomen. Kanske tolkas mätning som själva sinnebilden för den ”kvantitativa” metoden. Mätningens idé kan härledas till Galilei, som vi enligt Singer (1981, sid. 163) är ”mer skyldiga än någon annan för vår världsuppfattning, uttryckt i ett samspel mellan beräkningsbara krafter och mätbara kroppar”. Enligt Berka (1992, sid. 181) lydde Galileis motto: ”To measure what is measurable and to try to render measurable what is not so as yet.” Savage & Ehrlich (1992) menar dock att teorier om kvantitetens natur kan spåras åtminstone tillbaka till antikens grekland. Historien om hur mätningens idé också överfördes till beteendevetenskapen är intressant, om än i vissa stycken ganska mörk (se till exempel Gould, 1981). En historisk analys av mätning som idé ligger utanför syftet med denna avhandling. Detsamma gäller ett historiskt perspektiv på bedömningar i skolan som mycket sent, och endast i liten omfattning, sammanfaller med mätningens historia.

Ur ett metodperspektiv är det dock intressant att fundera litet över det modelltänkande som i allmänhet följer med kvantitativa data. Figur 3 visar en förenklad bild av matematisk modellering.



Figur 3 Förenklad bild av matematisk modellering.

Utvecklingen under 1600-talet, inte minst genom Galilei, innebar att vetenskapen anammade ett sådant modelltänkande. Genom att resultat från vetenskapliga undersökningar gavs en kvantitativ karaktär kunde de infogas i matematisk teori som utvecklades under samma period.

I användningen av modeller, kanske i synnerhet komplexa sådana, finns alltid en risk att modellen blir något mer än en modell.

> The unfortunate tendency to reify statistical models – to forget that they are descriptive summaries, not literal accounts of social processes – can only serve to discredit quantitative data analysis in the social sciences. (Gould, 1981, sid. 5)

I allmänhet beskriver statistisk dataanalys utfallen av sociala processer och inte processerna själva. ”It is therefore important to attend to the descriptive accuracy of statistical models, but to refrain from reifying them.” (Gould, 1981, sid. 6). Ett bra exempel på detta är klassisk testteori (se artikel I). Enligt mitt sätt att se beskriver klassisk testteori sambandet mellan provpoäng och ett tänkt verkligt mått på en individs prestationsnivå. Denna enkla modell är användbar för att kunna strukturera tänkandet kring osäkerheten i bedömningar. Den ska inte tolkas som en beskrivning av verkligheten så att det till exempel skulle kunna finnas en verklig motsvarighet till den sanna provpoäng som modellen talar om.

Sammanfattningsvis menar jag att det inte finns någon anledning av dra gränser mellan olika metoder, eller att generellt förorda en metod på bekostnad av andra. Olika metoder har olika förtjänster i olika sammanhang. Vilken metod som används beror i viss mån på en studies syfte, men en enskild metod ger alltid en sämre bild av en komplex verklighet än en mångfald av metoder.

Sammanfattning av mina utgångspunkter

Jag menar att de pedagogiska fenomen som är värda att studeras är komplexa och mångtydiga. De låter sig inte enkelt eller entydigt beskrivas och förstås eftersom de inte är entydiga och enkla, utan snarare mångfasetterade, komplexa och instabila. Detta innebär att olika perspektiv och metoder kan ge olika bidrag till en bild av de studerade fenomenen, och att det till och med det krävs en mångfald för att kunskapen om fenomenen ska bli så rik som möjligt. Jag menar vidare att vi i vetenskapliga sammanhang kan anta att det finns en verklighet, men att vi inte behöver göra det. De modeller och teorier som skapas är alltid konstruktioner, dvs. de kan användas för att på

ett mer eller mindre användbart sätt beskriva, förstå och förutsäga den verklighet som eventuellt finns där ute.

Den forskning som de fyra artiklar som ingår i denna avhandling representerar kan betecknas som problembaserad med en variation i vetenskapsteoretiska perspektiv. Som jag skriver i artikel IV så menar jag att väl genomförd forskning bidrar till vår kunskap om världen, nästan oavsett vilket vetenskapsteoretiskt eller metodologiskt perspektiv som anläggs. Denna problembaserade och perspektivrika forskning (observera att motsatsen till perspektivrik inte är perspektivfattig, utan snarare perspektivtrogen) förordar en eklekticism i såväl ansats som metod:

> Rather than aligning ourselves with one or another school of thought, researchers would do well to focus on significant pedagogical problems that call for solution, drawing on the concepts, theories, and methods of whatever disciplinary sources give promise of helping us deal with these problems more adequately. (Bellack, 1978, sid. 37, citerad i Lindblad, Linde, & Naeslund, 1999).

Enligt mitt sätt att se är det inte bara rimligt att använda sig av olika perspektiv och ansatser, oavsett vilken ”school of thought” de tillhör, utan det är till och med nödvändigt. En allsidig belysning av ett pedagogiskt problem kräver flera olika perspektiv. Detta synsätt överensstämmer också med Messicks utgångspunkter för validering: ”Hence, test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated” (Messick, 1989, sid. 14). Det är dock viktigt att vara medveten om sina utgångspunkter eftersom perspektivet möjliggör och sätter gränser för olika slutsatser i en empirisk studie. Oavsett vilket perspektiv som anläggs så bör det dock ske så medvetet och konsekvent som möjligt.

Några avslutande reflektioner

Frågor som handlar om bedömning i skolan har flera problematiska drag och kännetecken. För det första kännetecknas området av en konflikt mellan företrädare för olika inriktningar, framförallt utifrån ”traditionella” kontra ”innovativa” bedömningsformer. Denna konflikt kan i hög grad härledas till motsättningar mellan olika sätt att se på mål och kunskap (Shepard, 2000). För det andra görs ofta en åtskillnad mellan bedömning och undervisning.

> The impetus for my development of an historical framework was the observation by Beth Graue (1993) that “assessment and instruction are often conceived as curiously separate in both time and purpose” (Shepard, 2000, sid. 1).

För det tredje finns en utbredd omedvetenhet när det gäller konsekvenser av olika sätt att bedöma. “I propose a model which is based on deliberate and well thought-through choices, rather, than is too often the case, custom and practice. (Brown, 1999)

Med den utgångspunkt för värdering av bedömningars kvalitet som presenterats här menar jag att alla dessa problem kan förstås och även lösas. Utan att på förhand tala om vilka bedömningsformer som är rätt eller fel kan varje bedömningsform värderas utifrån sitt syfte. Det innebär bland annat att faktorer som kostnader i lärartid och pengar måste vägas mot samstämmighet med kunskapssyn och läroplanens mål. Genom att konsekvenser av bedömning också värderas, samt att bedömning betraktas ytterst som ett läromedel bland andra, så vävs allting ihop till en helhet. Genom en utvecklad valideringsmodell blir det möjligt att utforma bedömningssituationer på ett väl genomtänkt sätt.

Det som bedöms i skolan är mycket komplext. Strävan efter rättvisa i de komplexa inferenser som görs i skolan, till exempel att en elev ska ha ett visst betyg, kräver en mångfald och variation i formerna för bedömning. Jag menar att hög kvalitet i bedömning förutsätter en bredd i bedömningar avseende formella och informella bedömningssituationer, summativa och formativa syften, lärarbedömning och självvärdering, redovisningsformer, modeller för att sammanfatta och bokföra resultat och inferenser, typer av bedömningssituationer och så vidare. Gipps (1994) skriver:

> Designing assessments with fitness for purpose in mind will mean that a range of assessments is used; the acceptance of a range of types of assessments from traditional formal written examination to teachers’ own assessments … will have valuable

> spin-offs in terms of fairness (a range of approaches allows pupils who are disadvantaged by one assessment to compensate on another provided that they carry comparable weight) and cost (performance assessment and portfolios are expensive in terms of pupil and teacher time, standardized tests are less so). (sid. 163)

För att fatta välgrundade beslut om elevers kunskap och kompetens är det dock inte bara viktigt med en mångfald av olika bedömningssituationer, utan det krävs också att de enskilda bedömningarna har hög kvalitet. Detta ställer krav på en väl utvecklad modell för att granska kvaliteten hos bedömningar, och jag menar att den utgångspunkt jag presenterat fyller den funktionen.

Den empiriska forskning som presenterats i de artiklar som ingår i denna avhandling ger naturligtvis bara några små bidrag till kunskapen om hur validering kan ske och vilka problem och möjligheter som den utgångspunkt för validering som presenterats här kan uppvisa. I varje artikel finns förslag till fortsatt forskning inom det problemområde som artikeln behandlar. I tillägg till dessa skulle jag vilja föreslå tre inriktningar för fortsatt forskning.

För det första menar jag att kopplingarna mellan kunskapssyn och bedömningar, eller mål och bedömningar, behöver studeras. Är det så att dessa inte hänger ihop i lärares praktik, eller är det så att den bedömningspraktik som finns i själva verket överensstämmer väl med den syn på kunskap, mål och bedömningars syfte som råder bland lärare?

För det andra finns det anledning att studera konsekvenser av bedömningar. Inte mycket är känt om hur lärares bedömningspraktik påverkar elevers självkänsla och till exempel deras val av studievägar. Konsekvenser av storskaliga bedömningar, som till exempel de svenska nationella proven, är mycket litet studerade. På detta område finns en rad förutfattade meningar, som att nationella prov styr undervisning på olika sätt, men inte mycket empiriska belägg.

För det tredje kan utgångspunkten för värdering av kvalitet i bedömningar utvecklas vidare. Validitetsbegreppet har gått från att gälla en egenskap hos bedömningssituationen, till exempel ett prov, till att gälla en egenskap hos användningen av provresultat, det vill säga det är provresultatets användning som kan anses vara valid eller inte. I Messicks validitetsbegrepp handlar validering om en granskning av bedömningsresultatets inferenser och konsekvenser. Jag menar att detta innebär en begränsning som gör att vissa aspekter av såväl inferenser som konsekvenser inte omfattas av modellen. Ett exempel är de nationella proven i Sverige. Det är inte bara provresul-

taten som har inferenser och konsekvenser, utan själva företeelsen att dessa prov finns har olika effekter. Ett syfte med dessa prov är att bidra till tolkning och implementering av mål och kriterier, och denna påverkan kan tänkas ske oavsett vilka resultat som eleverna får på provet. Det faktum att proven finns och att lärare och andra har tillgång till dem har eventuellt en effekt på hur mål och kriterier tolkas. På motsvarande sätt har proven i sig sannolikt vissa konsekvenser för till exempel elever, oavsett hur de presterar på provet. Den kunskaps- och ämnessyn som förmedlas i provet kan påverka alla elever och det faktum att det finns nationella prov kan påverka elevernas sätt att förhålla sig till skolarbetet. Det finns alltså anledning att utveckla Messicks modell så att det omfattar inte bara inferenser och konsekvenser av resultat utan också inferenser och konsekvenser av bedömningssituationer och system.

Avslutningsvis vill jag framhålla att bedömningar är viktiga och centrala aktiviteter i utbildning och lärande. Patricia Broadfoot (2002) menar att det inte finns något tvivel om att bedömning i skolan är något mycket kraftfullt och inflytelserikt. Bedömningen har kapaciteten att hjälpa eller hindra, att vilseleda eller klargöra, att begränsa eller bemyndiga. Hon menar samtidigt att de ”teknologier” för bedömning som står till buds inte motsvarar de höga syften som ofta sammanknippas med bedömningsaktiviteter. Det finns många källor till osäkerhet och fel som blockerar vägen mellan insamling av belägg och olika inferenser. “Assessment results need to be treated with caution.” (Broadfoot, 2002, sid. 288). Kombinationen av att vara ett kraftfullt verktyg och samtidigt behäftat med stora svårigheter gör det än mer angeläget att hitta bra modeller för att värdera kvaliteten i olika bedömningar. I denna avhandling har flera kriterier för kvaliteter i bedömningar, ursprungligen presenterade som alternativ till validitet och reliabilitet, införlivats i och kombinerats med Messicks validitetsbegrepp.

Jag menar att Messicks validitetsbegrepp är en bra modell som konkretiserar centrala valideringsfrågor som kan användas som utgångspunkt vid en praktisk validering. Jag har försökt ge ett bidrag till att begripliggöra Messicks modell och även koppla den till det praktiska valideringsarbetet genom att ge exempel på relevanta frågor i förhållande till olika aspekter av validitet och även genom att påpeka hur bedömningens syfte samt läroplanens mål och kunskapssyn är nödvändiga för valideringen. En sådan utgångspunkt ökar möjligheterna att bli ”rätt mätt på prov” i den mest positiva tolkningen av detta uttryck.

English summary

The thesis consists of four papers and a text that summarizes the papers and presents a discussion of validation of educational assessments. A framework for validation is presented, that emanates from validity theory. It is argued that assessment purposes, epistemology, and curricular goals are necessary starting points for the validation.

Assessment in education is first of all a research field for studying all the activities in school-contexts that are used for the evaluation of performance in relation to curricular goals. The field has a broad sociological and pedagogical perspective, but also includes the methods and theories of educational measurement and psychometri.

Assessment in education is also a field of practice. Education refers in this context to all kinds of formal schooling guided by a curriculum. Assessment can be viewed both in a narrow and a broader sense. A narrow interpretation of assessment focuses on the direct evaluation of a performance, for example the scoring of a test. A broader interpretation sees assessment as a complex process including an assessment situation (a task and a context or format), a performance (related to the assessment situation), a result (a primary judgement like scores), inferences (a secondary judgement or an action), and consequences (value implications and social consequences of the assessment). In relation to this process, to validate means that every step in the process is scrutinized for possible flaws.

Validity

The concept of validity has for a long time been central to the development and evaluation of assessments. The concept is, however, not static, it has developed over the years, and in 1989 Messick (1989, p. 13) presented a very inclusive definition of validity:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.

In my work I have found his structuring of the validity concept both meaningful and useful. In Messick's view, validity is a unified concept. Despite this, for the sake of structure, he constructed a way of "cutting" the unified validity concept (Table 2).

Table 2 Messick's validity framework (Messick, 1989).

	Test interpre-tation	Test use
Evidential basis	**1**	**2**
Consequential basis	**3**	**4**

The rows of Table 2 represent the source of justification of the assessment, being either an evidential or a consequential basis. Assessments typically function as an evidential basis for inferences and actions, and to investigate if the evidence is good enough for the purpose of the assessment is an important part of validation. In addition to supplying evidence, assessments have consequences that should be justified in the validation process. In state mandated large-scale assessment programs, for example, the assessments are likely to influence the content and competences that are considered important. The columns of Table 2 represent the outcome of the assessment. Assessment results must always be interpreted, and this interpretation renders an inference about the group or individual participating in the assessment. In addition, assessment results are used as a basis for actions by those who arrange the assessment or by those who are assessed. The four fields of Table 2 point to the areas where arguments for validity of an assessment should be directed. To validate is to cover each cell (1-4) with theoretical arguments and empirical results.

A number of other approaches to quality in assessments have been proposed in the literature. During the 1990:s, several criteria for good assessments were reported as alternatives to validity and reliability as criteria for quality in assessment, for example by Linn & Herman (1997). In my view, these contributions fit very well in Messick's framework. In order to be useful in practice, meaningful questions have to be asked in relation to the theoretical framework. The combination of meaningful criteria and Messick's validity concept can take theory into practice. In this thesis, validation is seen as the process of acquiring and valuing empirical evidence as well as theoretical rationales that can support or question the quality of an assessment. I claim that this process has to build on perceptions of assessment purposes, curricular goals, and epistemology of education.

Assessment purposes

Different opinions have been presented concerning the possibility of using an assessment for multiple purposes. I tend to agree with those who claim that this often is a necessity primarily for efficiency reasons (e.g. Black & Wiliam, 1998). There is, however, no doubt that if an assessment has more than one purpose, each purpose must be validated separately. In fact, the purpose is central to the validation process because it summarizes the intention with the assessment, and evaluating outcomes in relation to intention is vital to validation. Even though, in my view, the general purpose of assessment in education must be learning, this is not a particularly useful categorization for the purpose of analysis. Several, more differentiated, lists of assessment purposes have been suggested (see e.g. Brown, 1999). The argument here is that the purpose of assessment must taken into account in the validation process

Curricular goals

The definition of education as formal schooling guided by a curriculum, will of course make curriculum important in the validation of the assessments which are intended to evaluate to what extent individuals and groups reach curricular goals. *Curriculum* is interpreted in different ways, and the corresponding term in Swedish ("läroplan") refers mostly to the written, intended curriculum. This is however not the only conceptualization of curriculum that is relevant for assessment and assessment validation.

Graybill (1998), for example, builds his discussion on three aspects of curriculum: the expressed, the implied, and the emergent. Curriculum can also be viewed as intended, implemented and experienced (Mullis et al., 2003). Assessment validation must not take only the intended, or expressed, curriculum into account.

Another aspect is the balance between general and more subject-specific goals in the curriculum. Learning goals can be viewed as subject specific and fairly narrow, focusing specific skills. Another view is that school-subjects are not primarily there for their own sake, but rather as good contexts supporting the aim of learning more general competences. For example mathematics could be taught not primarily because the concrete mathematical methods are important but because mathematics is a good context for the learning of structured problem-solving. The conceived balance between gen-

eral and specific goals is an aspect of curriculum that should influence teaching, learning and assessment, and therefore validation.

Epistemology

Epistemological issues are fundamental for assessment. A more elaborate discussion of epistemological aspects of assessment is far beyond the scope of this thesis. I only stress that there are three questions that need to be considered to some extent in the development and validation of assessments. The first question asks what knowledge and learning is. The assessor's answer to that affects assessment because it is reasonable that the assessment reflects the kind of knowledge that is intended. The second question concerns how qualities or levels of competence are conceptualized. Assessments will hopefully signal to the assessed the characteristics of different qualities, and in order to validate these consequences of assessment the assessor has to reflect on his or her views on these matters. The third question deals with the problem of how competence can be assessed. How is it possible to make judgements of the competence or understanding of others? Answers to this question can affect the way assessments are designed.

Final remarks

The empirical studies presented in this thesis show a variation of approaches and methods. Some of the studies can be characterized as post-positivistic, and one has a discourse-analytic approach. The methods used are "quantitative" as well as "qualitative".

In my view, the educational problems that are worth studying are complex and unstable. This means that in order to find some answers we need a variation in approach and method. In addition, the tension between "quantitative" and "qualitative" methods must be reconciled. In fact, this categorization is not particularly meaningful since for example, the construction of a questionnaire using selected-response items, often categorized as "quantitative method", requires very qualitative work for the formulation and selection of items and answers.

A number of different suggestions for further research are presented in the thesis. Each of the papers contains proposals for research directions with respect to the problems and areas of investigation. One of the suggestions for research on validation is to expand Messick's validity theory to include not only inferences from and consequences of assessment results but also infer-

ences from and consequences of the assessment situation. Particularly for large-scale assessment, the mere existence of national assessment programs has implications which could be viewed as a validity problem.

Finally, there are tendencies in Sweden as well as in other countries, that assessment is only seen as a formal activity that is external to teaching and learning. Research has shown that assessment can have significant influence on learning, and it is vital that assessment is viewed as part of the mission to improve learning. Assessments are powerful and influential, but the technologies for assessing are still lagging behind the aspirations for assessment (Broadfoot, 2002). Sources of inaccuracy and error stand between the collection of evidence and the formation of a judgement, and consequences of assessments are unintended and unwanted. This calls for greater awareness of assessment quality, and a framework for validation can hopefully contribute.

Summary of the papers

Paper I

Nyström, P. (in print). Reliability of educational assessments – The case of classification accuracy. *Scandinavian Journal of Educational Research*.

Reliability is a problem inherent in all educational assessments. This study focuses the problem of reliability of criterion-referenced assessments, in the context of Messick's (1989) validity concept. From this point of view, reliability is an aspect of validity. I argue that classification accuracy can be a meaningful way of conceptualizing reliability for a large number of educational assessments. Classification accuracy is defined as the extent to which the actual classifications of test takers agree with those that would be made on the basis of their true performances, i.e. the fraction of test takers who are subject to true-negative and true-positive classifications (Livingston & Lewis, 1995).

The paper presents a review of methods for the estimation of classification accuracy, and two methods are applied for the estimation of classification accuracy for two versions of a Swedish national test in mathematics. One method is described by Livingston & Lewis (1995). The other is a fairly transparent method, based on classical test theory, which is developed and presented in the paper. The two versions of a test actually refers to one test

administration where two different models for standard-setting are applied to the same score distribution. The study is based on the results of 1,201 students participating in a Swedish national test in mathematics. One model uses a cut-score for the total score on the test. This model is expected to strengthen reliability, since the judgement is based on a fairly large number of items. The other model uses three cut-scores, one for each content area in the test. This model strengthens alignment since the attainment can be connected to each domain, and students' mastery can be secured in several aspects. Separate cut-score can, on the other hand, lower reliability since each judgement is based on a smaller number of items.

The results from the simulation indicate that the classification accuracy is indeed lowered, from 0.93 to somewhere in the interval 0.89-0.92. The results suggest that there can be a significant trade-off between alignment and reliability when the scheme for performance standards is changed. As an aspect of validity, reliability competes with other aspects and changes of assessments intended to increase validity from one aspect can reduce validity from other aspects.

Paper II

Nyström, P. (2004). *National tests as a means of evaluating effects of streaming in Swedish upper-secondary school mathematics*. Submitted for publication.

This paper presents a study of attainment differences in mathematics related to the grouping of students into different study programmes in the Swedish upper-secondary school. For a sample of 403 students, results from two consecutive national tests in mathematics were compared, one at the end of compulsory school and the other at the end of the first mathematics course in upper-secondary school.

Even though Swedish upper-secondary school students are differentiated on 17 study programmes, the first mathematics course in the Swedish upper-secondary school is common for all students. Since almost every sixteen-year-olds continue from compulsory to upper-secondary school in Sweden, it is meaningful to compare attainment between these school-forms.

The results indicate that attainment is, on average, positively affected if students are grouped together with higher-achieving peers. This effect appears to be strongest for low-achievers. These results are consistent with

previous research. Some of the ways in which the streaming of students might cause these attainment differences are discussed in the paper. Furthermore, benefits and problems associated with using national tests for the evaluation of systemic effects are discussed. One problem is that the tests are not designed for this purpose, which limits the possible inferences from longitudinal studies. Another problem is that a more sophisticated analysis requires detailed records of performance, for example data on item level. If teachers are supposed to report data, this might take too much teacher time.

Paper III

Nyström, P. (2004). *Students' cues to perceived competence in mathematics.* Submitted for publication.

Paper III presents a short review of research that has pointed to the importance of expectancy beliefs in relation to learning, achievement and affect. Students' perceptions of their competence play an important role for motivational factors in school. The study addresses questions concerning how perceptions of competence are formed in school mathematics. How important are different indicators for students' views of their own competence in mathematics, and for conclusions as to the mathematical competence of their peers? The study is based on a questionnaire distributed to a sample of Swedish upper-secondary school students (n = 550).

The first research question examined the importance ascribed by students to some of the sources of information that they use for the formation of their own perceived competence. Based on the empirical results, three categories of cues to perceived competence were identified. The category that the students rate as most important can be characterized as an inner feeling of understanding. The category rated second most important consists of external cues, in this case results of formal educational assessments and judgements given by teachers. The least important category contains cues that are based on classroom behaviour, e.g. the number of question the student has to ask.

The second research question posed in this study concerned how students value the importance of different signs of success in mathematics? What characterizes the successful mathematics student? The importance that students ascribe to different signs of success in mathematics shows the same pattern as the importance given to different cues to perceived competence. The sign that the students found most important was that competent peers

understand how things are connected, even though they do not always remember everything. Similarly, the least important characteristic was that they know all the rules, even if they do not understand them. Similar to the findings concerning cues to their own perceived competence, students expect that to a high degree, successful students in general should understand what they are doing, and not just be able to do it.

The third research question investigated the differences in perceived competence between male and female students, and more specifically if such differences could be explained by the use of different cues to perceived competence in these groups. The results of this study indicate that the difference in perceived competence between male and female students cannot be explained by differential use of the cues investigated here. Meece et al. (1990) found that the links between efficacy-related perceptions and subsequent performance and enrolment patterns were similar for boys and girls. The results of this study are similar because male and female students seem to value cues to perceived competence in similar ways.

It is particularly interesting to consider the relative importance that students ascribe to test results as cues to perceived competence since the questionnaire was answered directly after the students had worked with a national test in mathematics for three hours. We therefore could have expected that the importance of test-results would have been rated higher than what we find in this study. This is relevant in a validation of the national test. Wiggins (1993, p. 147) asserts that : ”If the idea of ”consequential validity” is to have any teeth to it, we must consider the most predictable consequence of all formal evaluation: the likelihood that all results will influence the learner’s self-image and aspiration.” With more elaborate feedback, the influence of assessment on students’ perceived competence would, most likely, be stronger. There is firm evidence that innovations in classroom formative assessment, designed to strengthen the frequent feedback that students receive about their learning, yield substantial learning gains (Black & Wiliam, 1998).

Paper IV

Nyström, P. (2003). Lika barn leka bäst? En gymnasielärardiskurs om nivågruppering I matematik [Birds of a feather flock together? A teacher discourse on ability grouping in upper-secondary school]. *Pedagogisk Forskning i Sverige, 8*(4), 225-246.

The problems with and effects of ability grouping in primary and lower-secondary education have been studied extensively. There is, however, not much work reported on ability grouping of students over the age of sixteen. The aim of this study is to describe and interpret how mathematics teachers at an upper-secondary school express their experiences and beliefs concerning ability grouping.

This article presents results from interviews with six mathematics teachers, all working at the same Swedish upper secondary school. The study is based on a discourse-analytic approach (Potter & Wetherell, 1987), and the intention was to explore and present the variation in what teachers at this particular school say about ability grouping. What they say is not treated as a manifestation of a consistent belief, but rather as a statement made in a given context and with a certain purpose. As a background and interpretative tool, research on ability grouping and teacher beliefs are briefly reviewed.

The results are structured around four themes. The first theme concerns teachers' views on student diversity and the relevance of this diversity to ability grouping. Teachers identify student differences from a number of aspects, and the most characteristic ones are previous knowledge, study-ambition; motivation, interest, diligence, and emotions. All of these are considered to have some relevance to ability grouping. Five out of six teachers instantly answered yes to the question: Is student diversity an obstacle to learning in school?

The second theme covers the advantages, disadvantages, prospects, obstacles and risks that teachers express concerning ability grouping. The teachers in this study express a number of reasons for their generally positive evaluation of the kind of ability grouping that has been practised in their school. The teachers argue that an ability-grouping could, in a cost-efficient way, benefit the increasing number of students not passing the courses, and also offer students the opportunity to study more mathematics. In addition ability grouping was seen as a way of coping with the large diversity of students, and the teachers felt that grouping students enabled them to better stimulate the high-achieving students.

The list of advantages and possibilities is definitely longer for groups of high-achieving students compared with groups of low-achievers. Disadvantages and problems are almost entirely connected to groups of low-achievers, for example the risk that being exclusively with others who find mathematics

just as difficult and/or uninteresting as they themselves do does not give students good opportunities to learn. The students "hold each other back". Teachers also speak of problems concerning the formation of groups and the risk of placing students in the wrong group. Despite the teachers' focus on problems with ability grouping where low-achievers are concerned, teachers are generally positive to its possibilities and positive effects on all students. What consequences the awareness of drawbacks and risks has for what teachers do in the classroom is beyond the scope of this study.

Within the third theme teachers views on how teaching differs, or should differ, between classrooms with students at differing ability levels were explored. Teachers talk about two features differentiating teaching a high-achieving group from teaching a low-achieving group. With low-achieving groups, whole-class teaching activities must be significantly shorter and expectations on student activity, focus on task and perseverance are much lower.

The fourth, and last, theme deals with how teachers picture other people's views (e.g. headmasters, students or the public) on ability grouping. Teachers say that students are generally positive to ability grouping. The teachers' opinions differ regarding the views of people higher up in the school hierarchy, but they seem to feel free to create ability groups if they want to.

In summary, what these teachers express concerning the advantages and disadvantages of ability grouping varies a great deal. There are, however, some things they more or less agree on. They have a basically positive view of the way in which ability grouping has been practised at their school. They identify problems with ability grouping for low-ability groups, primarily related to the mixing of students who have trouble learning with those who primarily have motivational problems.

My interpretation is that the teachers try to find rational solutions to problems created by conflicting goals and external limitations. In their experience some students need more mathematics than the normal study programme offers, and some students do not learn enough mathematics to pass the courses. The teachers feel obliged to do something about this situation and ability grouping is considered to be a solution (or perhaps *the* solution), in particular given the restraints concerning group size and rooms that the teachers mention and their beliefs concerning mathematics and mathematics teaching and learning.

An intriguing result of this study is the relatively small differences that characterise teachers' descriptions of teaching groups with different achievement levels. The study presented here is not a classroom-study, it is a study of teacher-beliefs and discourse, but the lack of variation in how teachers describe classroom practice can be interpreted as an indication of what the practice looks like. At the very least it indicates that these teachers do not have a model for teaching mathematics that is varied enough to take advantage of the prospective possibilities created by the grouping of students.

Referenser

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). New York: McGraw-Hill.

Andersson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational outcomes*. New York: Longman.

Baker, E. L., Linn, R. L., & Herman, J. L. (1996). CRESST: A continuing mission to improve educational assessment. *CRESST Newsletter*.

Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: the engine of systemic curricular reform? *Journal of curriculum studies, 32*(5), 623-650.

Berka, K. (1992). Are there objective grounds for measurement procedures? I C. W. Savage & P. Ehrlich (Red.), *Philosophical and foundational issues in measurement theory* (sid. 181-194). Hillsdale, New Jersey: Lawrence Erlbaum.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning. The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Björkqvist, O. (1993). Social konstruktivism som grund för matematikundervisning. *Nordisk Matematikkdidaktikk, 1*(1), 8-17.

Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessment and testing*. London: RoutledgeFalmer.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-73.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The cognitive domain*. New York: McKay.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 25-27.

Broadfoot, P. (2002). Editorial. Beware the consequences of assessment. *Assessment in Education, 9*(3), 285-288.

Brown, S. (1999, August). *Fit for purpose assessment.* Paper presented at the EARLI-conference, Göteborg.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gilmore, A. (2002). Large-scale assessment and teachers' assessment capacity: learning opportunities for teachers in the National Education Monitoring Project in New Zealand. *Assessment in Education, 9*(3), 343-361.

Gipps, C. V. (1994). *Beyond testing. Towards a theory of educational assessment*. London: The Falmer Press.

Gould, S. J. (1981). *Den felmätta människan*. Stockholm: ALBA.

Graybill, E. (1998). Hermeneutics and education: The case for an emergent curriculum. *Curriculum and Teaching, 13*(1), 1-11.
Guba, E., & Lincoln, Y. (1994). Competing paradigms in qualitative research. I N. Denzin & Y. Lincoln (Red.), *Handbook of Qualitative Research*. Thousand Oakes: Sage.
Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education, 17*, 3-29.
Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 5-17.
Lindblad, S., Linde, G., & Naeslund, L. (1999). Ramfaktorteori och praktiskt förnuft. *Pedagogisk Forskning i Sverige, 4*(1), 93-109.
Linde, G. (2003). *Kunskap och betyg*. Lund: Studentlitteratur.
Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.
Linn, R. L., & Herman, J. L. (1997). *Standards-led assessment: Technical and policy issues in measuring school and student progress* (Nr. CSE Technical Report 426). Los Angeles: National Center for Research on Evaluation, Standards, and Student testing (CRESST), University of California.
Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197.
Lundgren, U. P. (1979). *Att organisera omvärlden. En introduktion till läroplansteori*. Stockholm: Liber.
Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*(1), 60-70.
Mehrens, W. A. (2002). Consequences of assessment: What is the evidence? I G. Tindal & T. M. Haladyna (Red.), *Large-scale assessment programs for all students. Validity, technical adequacy and implementation* (sid. 149-177). Mahwah, New Jersey: Lawrence Erlbaum Associates.
Messick, S. (1989). Validity. I R. L. Linn (Ed.), *Educational Measurement* (Vol. 3, sid. 13-103). New York: American Council on Education.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2003). *TIMSS Assessment frameworks and specifications 2003*. Boston: International Study Center, Lynch School of Education, Boston College.

Nyström, P. (1998). *Bedömning av kvalitet i matematikkunskaper* (PM Nr. 141). Umeå: Enheten för pedagogiska mätningar, Umeå universitet.

Nyström, P., & Palm, T. (2001a). Muntlig kommunikation och självvärdering. *Nämnaren, 28*(2), 36-40.

Nyström, P., & Palm, T. (2001b). Är det något fel med vanliga matteprov? *Nämnaren, 28*(1), 41-47.

Palm, T. (2001). *Performance and authentic assessment, realistic and real life tasks: A conceptual analysis of the literature* (EM Nr. 39). Umeå: Department of Educational Measurement.

Popham, W. J. (1997). Consequential validity: Right concern - wrong answer. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.

Romberg, T. A. (1992). Problematic features of the school mathematics curriculum. I P. Jackson (Ed.), *Handbook of research on curriculum*. New York: MacMillan.

Savage, C. W., & Ehrlich, P. (1992). A brief introduction to measurement theory and to the essays. I C. W. Savage & P. Ehrlich (Red.), *Philosophical and foundational issues in measurement theory* (sid. 1-14). Hillsdale, New Jersey: Lawrence Erlbaum.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

Shepard, L. A. (2000). The role of assessment in a learning culture. Presidential Address presented at the annual meeting of the American Educational Research Association, New Orleans, April 26, 2000.

Singer, C. S. (1981). *Naturvetenskapens historia*. Lund: Liber Läromedel.

Sivesind, K., Bachman, K., & Afsar, A. (2003). *Nordiske læreplaner*. Oslo: Læringssenteret.

Skolverket. (1996). *Grundskola för bildning: Kommentarer till läroplan, kursplaner och betygskriterier*. Stockholm: Skolverket.

Skolverket. (2000). *Naturvetenskapsprogrammet. Programmål, kursplaner, betygskriterier och kommentarer* (2000:14). Stockholm: Fritzes.

Skolverket. (2002). *Gymnasieskolans kursprov läsåret 2001/2002. En resultatredovisning*. Stockholm: Fritzes.

Skolverket. (2003). *Det nationella provsystemet - vad, varför och varthän?* (Dnr 01-2003:2038). Stockholm: Skolverket.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. I G. Tindal & T. M. Haladyna (Red.), *Large-scale assessment programs for all students. Validity, technical adequacy and imple-*

mentation (sid. 149-177). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Utbildningsdepartementet. (1994). *Läroplaner för det obligatoriska skolväsendet och de frivilliga skolformerna*. Stockholm: Utbildningsdepartementet.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph Nr. 6). Madison: National Institute for Science Education.

Wedman, I. (1988). *Prov och provkonstruktion*. Stockholm: Utbildningsförlaget.

Wiggins, G. P. (1993). *Assessing student performance. Exploring the purpose and limits of testing*. San Fransisco: Jossey-Bass Publishers.

Wolming, S. (1998). Validitet. Ett traditionellt begrepp i modern tillämpning. *Pedagogisk Forskning i Sverige, 3*(2), 81-103.

Åsberg, R. (2001). Det finns inga kvalitativa metoder - och inga kvantitativa heller för den delen: Det kvalitativa-kvantitativa argumentets missvisande retorik. *Pedagogisk Forskning i Sverige, 6*(4), 270-292.