Clearinghouse Forskningsserien - 2009 nummer 04

Pædagogisk brug af test

Et systematisk review

Teknisk rapport

af

Sven Erik Nordenbo
Peter Allerup
Hanne Leth Andersen
Jens Dolin
Helena Korp
Michael Søgaard Larsen
Rolf Vegar Olsen
Majken Mosegaard Svendsen
Neriman Tiftikçi
Rikke Eline Wendt
Susan Østergaard

Dansk Clearinghouse for Uddannelsesforskning

DANMARKS PÆDAGOGISKE UNIVERSITETSSKOLE
AARHUS UNIVERSITET

København 2009

Clearinghouse – Forskningsserien 2009 nummer 04

Pædagogisk brug af test – Et systematisk review

Dansk Clearinghouse for Uddannelsesforskning
er en enhed ved DPU, Aarhus Universitet

Titel	*Pædagogisk brug af test*
	- Et systematisk review
Copyright	© 2009 by Danish Clearinghouse for Educational Research
ISBN	87-7934-588-3
ISSN	1904-52-55
Sektion	**Clearinghouse - Forskningsserien 2009 nummer 04**
	Teknisk rapport
Reviewgruppe	Professor Peter Allerup, Aarhus Universitet,
	Professor Hanne Leth Andersen, Aarhus Universitet
	Institutleder Jens Dolin, Københavns Universitet
	Lektor, fil.dr. Helena Korp, Högskolan Väst
	Postdoc, dr.polit. Rolf Vegar Olsen, Universitetet i Oslo
Clearinghouse	*Dansk Clearinghouse for Uddannelsesforskning, Aarhus Universitet* ved:
	Professor Sven Erik Nordenbo, leder
	Lektor Michael Søgaard Larsen, souschef
	Kommunikationskonsulent Mette Thornval
	Videnskabelige assistenter:
	Majken Mosegaard Svendsen
	Neriman Tiftikçi
	Rikke Eline Wendt
	Susan Østergaard
	Studentermedhjælpere:
	Karen Agnild
	Jesper Larsen
	Lea Lund Larsen
	Jenny Vogel
	Ida Elbæk Wraae
Referencenummer	Projekt nr. 7500701
Publikationsmåned og -år	Maj, 2009
Denne rapport citeres som	Nordenbo, S.E., Allerup, P., Andersen, H.L., Dolin, J., Korp, H., Søgaard Larsen, M., Olsen, R.V., Svendsen, M.M., Tiftikçi, N., Wendt, R.E., & Østergaard, S. (2009) *Pædagogisk brug af test - Et systematisk review* . I: Evidensbasen. København: Dansk Clearinghouse for Uddannelsesforskning, DPU, Aarhus Universitet.

Kontaktadresse	Dansk Clearinghouse for Uddannelsesforskning
(adresse, telefon, e-mail)	DPU, Aarhus Universitet
	Tuborgvej 164
	2400 København NV
	Telefon: 8888 9980
	sen@dpu.dk
	www.dpu.dk/clearinghouse

Forord

Hermed udsendes det fjerde systematiske forskningsreview i forskningsserien, der er gennemført ved *Dansk Clearinghouse for Uddannelsesforskning*. Projektet har *Danmarks Pædagogiske Universitetsskole, Aarhus Universitet*, som opdragsgiver (jf. projekt, nr. 7500701) og er udført i perioden 23.5.2007 - 20.04.2009.

Det foreliggende arbejde er en *Teknisk rapport*, der redegør for alle dele af arbejdsprocessen. Reviewgruppen har deltaget ved datauddragning og som peer review gruppe. Den tekniske rapport er udarbejdet i et samarbejde mellem reviewgruppen og medarbejdere ved *Dansk Clearinghouse for Uddannelsesforskning*.

Clearinghouse ønsker at udtrykke sin tak til reviewgruppen, der beredtvilligt stillede sig til rådighed som deltagere i en uddannelsesforskningsopgave af en type, som ikke tidligere havde været gennemført i Danmark. Reviewgruppen takkes tillige for et godt samarbejde i projektets forskellige faser.

Clearinghouse ønsker også at takke *Danmarks Pædagogisk Bibliotek* for en forbilledlig hjælp og betjening med fremskaffelse af de mange dokumenter, som undersøgelsen bygger på.

Endelig ønsker Clearinghouse at takke *Danmarks Pædagogiske Universitetsskole* ved dekan Lars Qvortrup, fhv. prodekan Bjarne Wahlgren og prodekan Hans Siggaard Jensen for at tage initiativ og give støtte til gennemførelsen af opgaven.

Sven Erik Nordenbo

København, 20. april 2009

Sammenfatning

Hvad ønsker vi at få at vide?

[]

Hvem ønsker at vide det og hvorfor?

[]

Hvad fandt vi frem til?

[]

Hvad er implikationerne?

[]

Hvordan kom vi frem til disse resultater?

[]

Hvor kan der findes mere information?

Undersøgelsen indgår i *Evidensbasen*, som Dansk Clearinghouse for Uddannelsesforskning har etableret. Her kan man også finde et link til det forskningsgrundlag, *Koncept-notatet*, som styrer forskningsprocessen i *Dansk Clearinghouse for Uddannelsesforskning*, se www.dpu.dk/clearinghouse.

Indhold

Tabeller

Figurer

1 Baggrund

1.1 Baggrund

I de tre skandinaviske lande er der i løbet af de sidste par årtier taget politisk initiativ til introduktion af nationale test. Inspirationen hertil er bl.a. kommet fra Storbritannien, hvor den konservative regering i 1980'rne gennemførte en grundlæggende omdannelse af det engelske og walisiske skolesystem baseret på en ny forståelse af uddannelsessystemet, (Nordenbo, 2008).[1] I dette afsnit beskrives udviklingen i de tre lande på dette område.

1.1.1 Danmark

Introduktion af nationale test

23. september 2005 indgik den danske regering, Socialdemokratiet og Dansk Folkeparti forlig om en ny dansk folkeskolelov. Et vigtigt element var indførelsen af obligatoriske, nationale it-baserede test ("Obligatoriske test i folkeskolen," 2006). Aftalen blev til en ny lov vedtaget i det danske folketing 30. marts 2006, jf. LBK nr. 1195, se især §§ 13 & 13a (Paragrafferne er uændrede i den seneste LBK nr. 1049 af 28/08/2007). Et konsortium med COWI A/S i spidsen gik i juli 2006 i gang med at udvikle de nye test.

Tabel 1.1 opregner, hvilke fag der afholdes test i. Opgaverne bliver udviklet af en opgavekomission af op til 6 medlemmer, der er erfarne folkeskolere i de fag, testen omfatter. De tre første test i skoleåret 2006/2007 var i *Matematik* 6. klasse, *Dansk/læsning* 8. klasse, og *Fysik/kemi* 8. klasse. En evaluering i efteråret 2007 viste iflg. Det danske Undervisningsministerium (UVM):

> at testsystemet grundlæggende var velvalgt, men at de opgaver, der var blevet udviklet til den første udgave af testene, ikke var gode nok (jf. "De nationale test," Verificeret 2009.03.28).

UVM valgte derfor at lade 2008 fungere som et øve- og prøveår til udvikling af nye opgaver. Det har været intentionen, at op mod 900 skoler og 60-70.000 elever har skullet afprøve testopgaver til alle ti obligatoriske test, samt de to frivillige test i dansk som andetsprog. I skoleåret 2008/09 vil de nationale test, iflg. UVM, blive lanceret i en pilotfase fra marts til maj 2009. Formålet er at sikre, at testene virker som pædagogisk værktøj for lærerne, og at høste erfaringer med testene, før alle skoler skal anvende dem som obligatoriske test. Planen er, at de ti nye test er klar til foråret 2009. Hvis alt går vel, vil de obligatoriske test blive gennemført fra 2010.

[1] Det er værd at notere, at den engelske undervisningsminister Ed Balls i efteråret 2008 meddelte, at man har afskaffet obligatoriske nationale test i England og Wales for de 14 årige elever, men bevaret dem for de 11 årige, jf. (Depart. f. Children et al., Verificeret 2008.10.14).

Fag \ Klassetrin	1.	2.	3.	4.	5.	6.	7.	8.	9.
Dansk/læsning		x		x		x		x	
Matematik			x			x			
Engelsk							x		
Geografi								x	
Biologi								x	
Fysik/kemi								x	
Dansk som andetsprog *)					x		x		

*) DE TO PRØVER I *DANSK SOM ANDETSPROG* ER FRIVILLIGE TEST

Tabel 1.1: Plan for afholdelse af nationale test i den danske folkeskole

Nye træk ved de nationale test

Den nye testpraksis adskiller sig på flere punkter fra den hidtidige i den danske folkeskole. Det danske undervisningsministerium (UVM) lister på deres hjemmesider følgende punkter, jf. ("Obligatoriske test i folkeskolen," 2006) og ("De nationale test," Verificeret 2009.03.28): Testene er it-baserede og gennemføres ved, at eleverne besvarer dem på en computer. De stilles gratis til rådighed for skolerne af UVM. Testene er selvscorende – lærerne skal ikke selv rette dem, men får leveret resultaterne. Og endelig er de adaptive, hvilket er det helt særlige ved disse test.

Der er tre træk ved de punkter, som UVM opregner, som i første omgang kræver nærmere kommentarer: nemlig at testene er it-baserede, at de er adaptive, og at de nye test er centralt besluttede og tilrettelagt.

Testene er it-baserede

At de nye test er it-baserede, beskriver UVM gennem et scenario, de ser som den nye virkelighed efter 1. marts 2007 (jf. "Obligatoriske test i folkeskolen," 2006):

> Når læreren har givet klassen adgang til at gå i gang, logger eleven sig på med sit personlige brugernavn og password. Herefter har eleven som udgangspunkt 45 minutter til at besvare opgaver i. I løbet af de 45 minutter vil eleven skulle svare på ca. 60 spørgsmål. Men læreren kan vælge at forlænge testen for elever, der måtte have behov for mere tid. For at kunne vurdere om eleverne skal bruge mere tid, kan læreren på sin computer følge med i, hvordan den enkelte elev klarer sig.
>
> Hver gang eleven har besvaret en opgave, går der besked til en central computer om elevens besvarelse og på grundlag af besvarelsen (rigtig/forkert) udtrækker og fremsender den centrale computer en ny opgave.
>
> Computeren henter opgaverne i en "opgavebank", som er en database med et stort antal opgaver, som er udviklet i løbet af efteråret 2006. For at sikre en me-

get høj kvalitet af opgaverne, er hver af opgaverne inden afprøvet på ca. 500 elever. Opgaver, der ikke lever op til skrappe kvalitetskrav, er kasseret.

Hvad er nu for det første begrundelserne for disse nationale, it-baserede test? Især to springer i øjnene som centrale, den ene begrundelse stærkt betonet, den anden noget mindre, nemlig henholdsvis pædagogiske hensyn og kvalitetssikringshensyn.

Det fremhæves, at disse test er et pædagogisk værktøj, der skal anvendes sammen med resultaterne af evalueringen til brug for den videre planlægning af undervisningen, i vejledningen af den enkelte elev, og i underretningen af forældrene med henblik på at tilrettelægge en undervisning og et forældresamarbejde, der understøtter eleven bedst muligt.

Men det nævnes også, at læreren, skolelederen og kommunalbestyrelsen på hver sin måde får adgang til testdata. Læreren får adgang til alle oplysningerne om både klassens og den enkelte elevs resultater. Yderligere får læreren adgang til at se de opgaver, der indgår i testforløbene, og hvordan eleven har besvaret hver enkelt opgave. Skolens leder, der har til opgave at forestå den pædagogisk ledelse af skolens lærere, får adgang til oplysninger på skole- og klasseniveau, samt et overordnet overblik over de enkelte elevers resultat. Lederen har endvidere til opgave at orientere skolebestyrelsen om resultaterne på skoleniveau. Kommunalbestyrelsen, der har til opgave at føre tilsyn med skolerne og den kommunale forvaltning, får adgang til testdata på kommune- og skoleniveau.

Det er endelig planen, at testdata opgøres og offentliggøres på landsniveau. Denne opgørelse vil ske både på det samlede testresultat og på profilområder. Hvert fags testspørgsmål opdeles i tre emnemæssige områder, de såkaldte profilområder, jf. Tabel 1.2. Resultaterne på profilområder betegnes samlet som en national præstationsprofil, og den anvendes bl.a. ved vurderingen af den individuelle elevs præstationsprofiler. Tilbagemeldingen til forældre og elever sker ved hjælp af en algoritme, der genererer beskrivelser i almindeligt sprog om den individuelle elevs præstationsprofiler baseret på den score, de har opnået i testen.

Fag/klasse	Profilområde 1	Profilområde 2	Profilområde 3
Matematik 6. klasse	Tal og algebra *)	Geometri	Matematik i anvendelse
Læsning (dansk) 8. klasse	Sprogforståelse	Afkodning*	Tekstforståelse
Fysik/kemi 8. klasse	Energi og energiomsætning	Fænomener, stoffer og materialer	Anvendelser og perspektiver
Biologi	Den levende organisme	Levende organismers samspil med hinanden og deres omgivelser	At bruge biologien: Biologiens anvendelse, tankegange og arbejdsmetoder
Geografi	Naturgrundlaget	Kulturgeografi	At bruge geografien
Dansk som andetsprog	Kommunikative færdigheder - det skrevne sprog	Sprog og sprogbrug	Viden om sprogtilegnelse og egen læring
Engelsk 7. klasse	Læsning	Ordforråd	Sprog og sprogbrug

*) PROFILOMRÅDERNE AFKODNING (LÆSNING) OG TAL OG ALGEBRA (MATEMATIK) INDGIK IKKE I TESTENE I 2007

Tabel 1.2: Profilområder i de nationale danske test

Ved beregning af præstationsprofilerne vil der ske en statistisk korrektion for forskelle i elevernes baggrund (nemlig med hensyn til køn, etnicitet, forældres uddannelsesbaggrund og indkomstforhold) på både skole- og kommuneniveau. Målet hermed er at foretage en vurdering af, hvordan den enkelte skole/kommune ville have klaret sig, hvis elevsammensætningen havde været som landsgennemsnittet. Som et supplement til de faktisk registrerede resultater vil den enkelte skole og kommune få adgang til disse korrigerede resultater, og de vil ligesom andre testdata skulle behandles fortroligt.

Der er altså tale om, at systemet medfører, at UVM kan foretage en detaljeret benchmarking på klasse-, lærer-, skole-, kommune-, regions- og landsniveau.

Testene er adaptive

For det andet er testen adaptiv. Det betyder, at den tilpasser sig de enkelte elevers færdighedsniveau undervejs i testforløbet. En elev vil normalt starte med en middelsvær opgave. Undervejs gælder det, at svarer eleven rigtigt på to på hinanden følgende opgaver, får eleven næste gang en sværere opgave, mens eleven ved to på hinanden følgende fejlsvar får en lettere opgave. Når svarene skifter mellem rigtigt og forkert, får eleven opgaver af samme sværhedsgrad, og man fortsætter, indtil der totalt set er besvaret et tilstrækkeligt antal opgaver, der sikrer præcision i bestemmelse af elevens færdighedsniveau. Dette princip gælder gennem hele testen.

Der er et pædagogisk argument bag anvendelsen af adaptive test. Ved en normal test vil de fleste elever oftest opleve, at der er opgaver, som både er for nemme, passende og for svære. Med test efter det adaptive princip vil den enkelte elev hovedsageligt få opgaver med et passende sværhedsniveau, uanset om eleven er meget dygtig eller

mindre stærk i faget. Ifølge UVM giver det erfaringsmæssigt et mere udfordrende og effektivt testforløb for den enkelte elev, når eleven ikke skal spilde tiden på opgaver, der er for lette eller lide nederlag ved at skulle besvare opgaver, der ligger ud over elevens faglige niveau. Hver elev får altså i praksis i rent sværhedsmæssig henseende sin egen skræddersyede test, og der vil næppe, mener UVM, være to elever, der får præcist den samme kombination af opgaver.

Forudsætningen for, at det pædagogiske motiv bag det adaptive princip kan gennemføres i praksis er, at opgaverne tilfredsstiller bestemte psykometriske principper. Testen skal statistisk set opbygges efter Rasch-modellen, se herom fx (Allerup, 1987), der bygger på arbejder, som den danske statistiker Georg Rasch udviklede i perioden 1952-1977. I dag er principperne diskuteret, videreudviklet og anvendt i international sammenhæng (Se fx "Institute for Objective Measurement,"). Disse principper tillader, når visse betingelser er opfyldt, at give en beskrivelse af en elevs individuelle færdighedsudvikling i en faglig disciplin uafhængig af, hvilke konkrete opgaver eleven er præsenteret for. Den adaptive test efter Rasch-modellen kan derfor ses som et effektivt redskab til at konstatere, hvor en elev befinder sig objektivt set i forhold til andre elever i en skoleårgang og set over flere år, om der sker passende fremskridt inden for det testede fagfelt. Netop udformningen af test efter Rasch-modellen giver derved det tekniske redskab til både at benchmarke og kvalitetsvurdere den danske folkeskole.

Testene er tilrettelagt centralt

Hertil kan endelig tilføjes det tredje punkt, nemlig at de nye test er centralt besluttede og tilrettelagt. Heraf følger: Det er op til den centrale administration at afgøre, *om* man skal have test, det er ikke op til den enkelte lærer, skole eller kommune. Det er op til den centrale administration at afgøre, *hvornår* man skal have test, det er ikke op til den enkelte lærer, skole eller kommune. Og det er op til den centrale administration at afgøre *indholdet* af hver enkelt test, det er ikke op til den enkelte lærer, skole eller kommune.

På denne baggrund er det klart, at det har været magtpåliggende for UVM, at får et sådant redskab i hænde, at det nu – i modsætning til tidligere - bliver muligt hurtigt og omfattende at danne sig et overblik over den faglige situation i de fag, de nationale test omfatter.

1.1.2 Norge

Introduktion af nationale test

Den uddannelsespolitiske baggrund for indførelse af de nationale test i Norge var to rapporter fra et udvalg nedsat af regeringen. Udvalget havde et bredt mandat (NOU, 2002, 2003), og dets udgangspunkt var, at der ikke eksisterede et nationalt system til at overvåge kvaliteten i undervisningen, noget som blandt andet blev påpeget i en OECDvurdering af norsk uddannelse (OECD, 1989). Udvalget foreslog blandt andet at indføre obligatoriske nationale test i nogle grundlæggende kompetencer på tværs af de konkrete skolefag. Testene skulle være indikatorer for kvaliteten af resultaterne. Videre foreslog udvalget, at denne information skulle implementeres i et elektronisk rapporteringsværktøj. Ideen var, at dette værktøj skulle give offentlig indsigt i forskellige

aspekter ved skolernes kvalitet. Skolernes resultater i de nationale test blev foreslået som indikatorer på resultatkvalitet. Andre indikatorer skulle give indblik i andre sider af skolernes kvalitet (proces- og strukturkvalitet). Dette forslag blev i grove træk vedtaget af Stortinget (UFD, 2002), og fulgt op i en stortingsmelding, *Kultur for læring (UFD, 2004)*. Det blev vedtaget at indføre nationale test i skrivning, læsning, matematik og engelsk i slutningen af 4. 7. og 10. klasse i grundskolen og ved slutningen af det første år i den videregående skole.

Der var en bred politisk tilslutning til de nationale test. Den politiske uenighed bestod primært i forskellige syn på, hvilken information som skulle offentliggøres.

Udvikling af de nationale test

Faggrupper fra forskellige universitetsmiljøer i landet fik til opgave at udvikle testene. Foråret 2004 blev de nationale test gennemført i læsning og matematik for 4. og 10. klasse og i engelsk, læsning og skrivning i 10. klasse. I foråret 2005 blev test gennemført på alle de fire vedtagne områder på alle klassetrin. De fleste test blev gennemført som skriftlige prøver, men prøven i engelsk og læsning var en elektronisk adaptiv test. Med undtagelse af engelsktesten rettede lærerne selv testene, og resultaterne blev offentliggjort i form af normerede skalaer på skoleniveau på den nye hjemmeside http://skoleporten.utdanningsdirektoratet.no.

Evaluering af de nationale test

Testene i 2004 og 2005 blev evalueret grundigt. Evalueringen af de faglige test bestod af både psykometriske evalueringer (Lie et al., 2004; Lie et al., 2005) og undersøgelser af brugernes erfaringer og vurderinger (Kavli et al., 2005; TNS-Gallup, 2004). Til sammen viste disse evalueringer, at det eksisterende testsystem havde en mangelfuld kvalitet på flere områder:

- Der eksisterede ikke tydelige krav til testenes indhold og form. Dette førte til, at de forskellige faggrupper havde udviklet test som var forskellige i form og indhold.
- Nogle test havde for svage psykometriske egenskaber til at blive publiceret på den nye hjemmeside.
- Det viste sig, at ganske mange elever boykottede testene i 10. klasse og på 1. klassetrin i videregående skole, og nogle af testene blev gjort offentlig tilgængelige før selve testen fandt sted.

Lie mfl. (2005) konkluderede, at testsystemet som helhed havde så store mangler, at man burde stoppe de nationale test i det efterfølgende år, så man kunne bruge tid til udvikle og konsolidere systemet.

Denne anbefaling blev fulgt og i efteråret 2007 var et nyt system for nationale test på plads. De viktigste ændringer fra det første forsøg opsummeres nedenfor:

- Tydelig forankring og specifikation af krav:
  - Nye læreplaner var indført, hvor de grundlæggende færdigheder (som de nationale prøverne har til hensigt at måle) var implementeret i målene for de en-

kelte fag. Dermed var de nationale test klarere forankret i målene for undervisningen.
  - Et overordnet nationalt kvalitetsvurderingssystem blev etableret, hvor hensigten med de nationale test er tydeligere afgrænset.
  - De faggrupper, der har til opgave at udvikle testene, har fået en tydeligere instruks. Det er nu testgrupperne selv, der skal dokumentere, at testene lever op til eksplicitte psykometriske standarder.
  - Hver enkelt faggruppe har på forhånd udviklet eksplicitte rammer, som redegør for den kompetence, som testene skal måle.

- Tydeliggørelse af at testene hovedsalig skal bruges som pædagogiske resurser:
  - Testene afholdes om efteråret i 5. klasse og i 8. klasse. Begrundelsen er, at testene skal være pædagogiske resurser, og de afholdes derfor på de tidspunkter, hvor eleverne er i en overgangsfase mellem hovedtrin i grunduddannelsen.
  - Resultater rapporteres i form af niveauer – tre niveauer for testene i 5. klasse og 5 niveauer for testene i 8. klasse. Der er blevet udviklet verbale beskrivelser af, hvad elever på de forskellige niveauer typisk behersker. Hvordan niveauerne og de tilhørende beskrivelser er blevet etableret, er imidlertid svagt dokumenteret. Der er brugt en eller anden variant af såkaldt "bookmark"-procedure, se for eksempel (Cizek & Bunch, 2007).
  - Det findes vejledningsmateriale både for lærere i de enkelte fag, for skoleledere og skoleejere som har til hensigt at vise, hvordan resultater af testene kan bruges i skolens/lærerens pædagogiske arbejde.
  - Offentlig adgang til resultaterne er reduceret for at forhindre at testene primært bliver brugt til at sammenligne skoler[2]. Kun resultater på kommuneniveau er offentlig tilgængelige på http://skoleporten.utdanningsdirektoratet.no.

Test i skrivning blev afviklet, fordi det ikke syntes at være muligt at udvikle test med tilstrækkelig psykometrisk kvalitet. Der blev også indført et nyt elektronisk indrapporteringssystem som skulle sikre dataindsamlingen. Tabel 1.1 giver en oversigt over de fagområder, der testes i.

[2] Dette har imidlertid vist sig vanskeligt, fordi aviser med støtte i andet lovstof har sikret sig ret til indsigt, noget som har ført til, at det fortsat er store opslag i nyhedsmedier hvor tabere og vindere udråbes.

Fagområde / Grundskolen	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Læsning	(x)	(x)			x			X		
Matematik		(x)			x			X		
Engelsk					x			X		

Tabel 1.3: Plan for afholdelse af nationale test i den norske skole

En brugerundersøgelse blandt lærere, skoleledere og forældre viser i store træk større andele af positive vurderinger fra alle brugergrupper. Skolelederne udtrykker, at de vil følge op på resultaterne fra de nationale test. Videre siger forældrene, at de er blevet informeret om resultaterne for deres barn. Men de fleste skoleledere og lærere rapporterer, at testene bidrager til at identificere faglige styrker og svagheder hos eleverne. Samtidig rapporterer de, at dette ikke er ny information om eleverne. Derudover er det administrative system, der skal registrere data, blevet kraftig kritiseret for at være tidskrævende (Kavli, 2008).

Udover de nationale test har man i løbet af de sidste par år etableret såkaldte kortlægningsprøver. Der findes i dag sådanne prøver i læsefærdighed i 1. og 2. klasse og talforståelse og regnefærdighed i 2. klasse. Disse prøver skiller sig ud fra de nationale test ved, at prøverne er meget lette for de fleste elever. Der er ikke er en national indrapportering af data, de bliver heller ikke brugt til at rapportere indikatorer på skoleniveau (selv om nogle kommuner, for eksempel Oslo, bruger prøverne til dette), og de samme prøver bruges flere år efter hinanden. Kortlægningsprøverne har imidlertid også mange fællestræk med typiske nationale testsystemer. Derfor vælger vi at inkludere dem i beskrivelsen af det norske nationale prøvesystem. Fællestrækkene er: De er obligatoriske, de drejer sig om centrale faglige færdigheder relateret til læsning og regning, der findes en rimelig standardiseret protokol for, hvordan prøverne skal administreres og prøverne rapporteres efter en skala som bruges til at etablere et såkaldt kritisk niveau. Formålet med kortlægningsprøverne er - tydeligere end de nationale prøver - ment som en formativ vurdering til at afdække behov for opfølgning og tilrettelæggelse på individ- og skoleniveau. (www.udir.no – 31. Mars 2009), specielt i forhold til at afdække hvilke elever som har behov for særskilt opfølgning tidligt i undervisningsforløbet. I tilgift udsendes et righoldigt vejledningsmateriale, som består af lærervejledning og forslag til, hvordan lærerne kan følge op på resultaterne i deres undervisningsaktiviteter.

1.1.3 Sverige

Sverige har anvendt forskellige former for nationale test i godt og vel tres år. De test vi i dag har, er koblet til den målstyrede læreplan og det kriterierelaterede karaktersystem som blev indført i Sverige i 1994. De nationale test er udformet som en integreret del i den målstyrede skole for at støtte implementeringen af læreplaner, fagplaner og vurderingen af elevernes kundskaber. Før 1994 var den svenske skole regelstyret og betydeligt mere centraliseret end i dag. Karaktersystemet var relativt, dvs. i karaktergivningen relateredes hver elevs præstationer i forhold til andre elevers præstationer og ikke til mål eller standarder. I dette system havde de nationale test som eneste funktion at

kalibrere karaktergivningen i overensstemmelse med normalfordelingen. Resultatet kunne derfor ikke, som i dagens system, anvendes pædagogisk som for eksempel til at tydeliggøre mål for uddannelsen eller vise elevernes stærke og svage sider. De kunne heller ikke anvendes for resultatopfølgning og styring, da de ikke gav information om elevernes kundskabsniveau.

I dag omfatter det nationale testsystem en mængde test og diagnosemateriale, hvor sigtet er, at de skal kunne anvendes fra førskoleklassen til og med gymnasiet. I testsystemet indgår også en national testbank, hvor lærere bliver tilbudt oplysninger og testmateriale i visse fag, hvor der ikke er nationale test.

Det nationale testsystems formål

- Testsystemet har flere formål:
  - at bidrage til øget målopfyldelse for eleverne
  - at tydeliggøre mål og vise elevernes stærke og svage sider
  - at konkretisere fagmål og kriterier for karakter
  - at støtte en ligeværdig og retfærdig bedømmelse og karaktergivning
  - at give grundlag for en analyse af i hvilken udstrækning kundskabsmål nås på skoleniveau, på regionalt og nationalt niveau.

I hvilke fag og år gennemføres nationale test?

I grundskolen gives nationale faglige test i svensk, svensk som andetsprog, matematik og engelsk i 5. og 9. klasse. Fra og med foråret 2009 gives også faglige test i svensk, svensk som andetsprog og matematik i 3. klasse og i biologi, fysik og kemi i 9. klasse, jf. Tabel 1.4.

Fag	Grundskolan 1.	2.	3.	4.	5.	6.	7.	8.	9.	Gymnasiet
Svensk			x		**x**				**x**	**x**
Svensk som andetsprog			x		**x**				**x**	
Matematik			x		**x**				**x**	**x**
Engelsk			x		**x**				**x**	**x**
Biologi									x	
Fysik									x	
Kemi									x	

TEST ANGIVET MED X INDFØRES I FORÅRET 2009

Tabel 1.4: Plan for afholdelse af nationale test i den svenske skole

I gymnasiet gives nationale faglige test i kernefagene svensk, engelsk og matematik. Disse kurser er fælles for alle gymnasieprogrammer, og testen udføres således af alle elever. I programmer med fordybelseskurser i engelsk og matematik findes yderligere emnetest for disse kurser.

I grundskolens år 9 og i gymnasiet skal resultatet fra de nationale test indgå i grundlaget for karaktergivningen.

At udvikle nationale test

Det er Skolverket, som på regeringens vegne, har ansvaret for de nationale test. Det konkrete arbejde med at udvikle testen gennemføres dog på forskellige universitetsinstitutioner i landet, hvor videnskabelige fag- og fagdidaktiske kompetencer findes. Således er der altid metodekompetencer for bearbejdning og analyse af resultater fra forsøg. Også aktive lærere involveres i forskellige faser af processen. I udviklingsarbejdet indgår blandt andet, at i relation til testens formål og anvendelse fastsættes en ramme for testens indhold, hvilket format som oplysningerne skal have, og hvordan testen skal gennemføres. Der skal også udarbejdes retningslinier for, hvordan oplysninger skal udvælges og testes i forhold til de mål, som de skal testes efter, hvordan bedømmelsen skal ske og hvilke principper som skal anvendes, når der for eksempel skal sættes karaktergrænser. Arbejdet følger den alment omfattede kvalitetsnorm for testfremstilling og testanvendelse *Standards for educational and psychological testing* som er udviklet fælles og fastsat af American Educational Research Association, American Psychological Association og National Council on Measurement in Education. At udvikle en national faglig test tager i Sverige mellem 1,5 og 2 år, hvilket er det samme som i andre lande (Skolverket, 2004).

De nationale test rettes af de lærere, som underviser eleverne. De nationale test i engelsk og matematik for 9. klassetrin er hemmeligstemplede i ti år. Det indebærer at nogle af oplysningerne kan genbruges. Årsagen til dette er dels økonomiske årsager, dels at det letter sammenligninger af elevernes kundskaber over tid. Testen for 5. klassetrin nykonstrueres hvert andet år.

Hvordan ser testen ud?

Grundskolens faglige test består af tre deltest, hvoraf én er mundtlig og kan gennemføres på et valgfrit tidspunkt. De øvrige deltest gennemføres på bestemte tidspunkter, som fastsættes af Skolverket. Også tiden som eleverne har til de forskellige deltest, er bestemt på forhånd. Testen kan tilpasses elever med behov for særlig støtte. Det kan handle om, at elever får mere tid eller får læst oplysningerne op for at kunne gennemføre testen. I visse dele af testen er hjælpemidler såsom lommeregner og leksikon tilladte. Det angives i de respektive deltest, hvilke hjælpemidler, der er tilladte.

De nationale emnetest er konstruerede, så de kan afdække forskellige aspekter af elevernes kunnen, men også mere komplekse evner hos eleven kan medtages. Dog kan ingen test vise elevernes kundskaber i forhold til samtlige mål i kursusplanen. Selvom det tilstræbes at inkludere autentiske og kreative oplysninger i testen, er der en begrænsning i testens format. Både af hensyn til formative og summative formål må de nationale test derfor komplementeres med andre oplysninger af varierende art.

Nedenfor beskrives de nationale test indhold og form i grundskolen, jf. (http://www.skolverket.se/sb/d/2519).

Engelsk. Testen i engelsk består af mundtlig interaktion og produktion, læse- og lytteforståelse samt skriftlig produktion. En deltest har fokus på at samtale og tale og gennemføres med to, alternativt tre til fire, elever. I den deltest, som fokuserer på læseforståelse forekommer længere og/eller kortere tekster med flere valgmuligheder eller spørgsmål med åbne svar, hvor eleverne selv skriver svaret med et eller flere ord. Matchningsoplysninger og forskellige typer af tekster med huller kan også forekomme. I deltesten som tester lytteforståelse forekommer for eksempel interview efterfulgt af spørgsmål med åbne svar og/eller flere valgmuligheder. En anden testtype kan være korte samtaler eller monologer med oplysninger af matchningskarakter. I deltesten, som støtter bedømmelsen af elevernes evne til at skrive på engelsk, tilbydes i reglen to emner af forskellig art.

Matematik. Testen i matematik giver eleverne mulighed at vise deres evner på forskellig vis. De forskellige dele adskiller sig, hvad gælder kundskaber om indholdet, arbejdsform, redegørelses- og vurderingsform. Emnetesten i matematik på grundskolens årskursus 9 omfatter fire dele: En mundtlig deltest som gennemføres i en gruppe, en deltest, hvor kun svar kræves og som gennemføres uden adgang til regnemaskine, en deltest med en større mængde oplysninger, som kræver en udførlig redegørelse, og endelig en deltest med oplysninger og hvor der skal redegøres udførligt for resultatet. De to sidste deltest gennemføres med adgang til regnemaskiner. I den del, hvor eleven ikke må anvende regnemaskiner kan eleven primært vise kvaliteter som begrebsforståelse, hovedregning, talopfattelse etc.

Svensk og svensk som andetsprog. Testen i svensk og svensk som andetsprog har altid et tema som karaktariserer de forskellige deltest. Til testen hører et teksthæfte med en blanding af skønlitterære og fagprosatekster. Emnetesten for grundskolens årskursus 9 består af tre dele. Den første del tester elevernes mundtlige evner og udgøres af test som eleverne lytter til og efterfølgende diskuterer. Den anden del tester elevernes læseforståelse ud fra teksthæftet. Den tredje deltest tester elevernes skriftlige evner og består af en skriftlig opgave med samme tema som teksthæftet. Kursustesten i svensk B og svensk som andetsprog B består af en mundtlig deltest, hvor eleverne laver en præ-

sentation af testens tema. En kortere skriftlig opgave, tester elevens læseforståelse og evne til kortfattet at præsentere tekster. I en længere skriftlig opgave producerer eleverne for eksempel en argumenterende, undersøgende eller foredragslignende tekst. Ved besvarelse af skriveopgaverne er det tilladt at anvende ordbog til og fra modersmålet.

Fremtiden

Både grundskolen og gymnasiet står for at skulle indføre nye læreplaner og nyt karaktersystem. Dette kommer til at påvirke de nationale test, men der er endnu ikke udarbejdet et direktiv til Skolverket om, hvordan det nationale testsystem skal tilpasses de nye vilkår og krav, som forandringerne forventes at medføre. Alt peger på, at forandringerne kommer at medføre flere nationale test. Ifølge regeringens forslag skal det nye karaktersystem stadig være mål- og kundskabsrelateret, men have flere trin end i dag.

Når det gælder grundskolen foreslås i Regeringens proposition 2008/09:87 *Tydligere mål og kunskapskrav - nya læreplaner for skolen,* at der udover de obligatoriske nationale emnetest i årskursus 3 og 9, skal der også være emnetest i svensk, svensk som andetsprog, matematik og engelsk i slutningen af årskursus 6. Ifølge propositionen bør der også være mulighed for nationalt at afstemme årskursus 3 for fag indenfor de naturorienterede og samfundsorienterede områder.

I betænkningen *Framtidsvägen - en reformerad gymnasieskola* (SOU 2008:27) foreslås, at antallet af gymnasieprogrammer reduceres, og at der kommer en tydeligere adskillelse mellem erhvervs- og studieforberedende programmer, hvor de erhvervsfaglige ikke giver kompetencer til videregående uddannelser. De nuværende kerneemner erstattes med såkaldte programfælles emner. Det betyder, at visse emner skal indgå i samtlige programmer, men have forskellig omfang og udformning. En konsekvens er, at de nationale test ikke kan være ens i alle programmerne. Forskerne forslår også, at flere nationale test indføres i gymnasiet.

1.1.4 Sammenfatning

[]

1.2 Problemfelt

Mens det næppe kan anfægtes, at de nationale test er udformet på en sådan måde, at det ene hovedformål om kvalitetskontrol og benchmarking vil kunne opfyldes, er det ikke i samme grad åbenbart ud fra den eksisterende forskning, om det andet hovedformål kan opfyldes, nemlig om og hvordan de oplysninger om de enkelte elevers faglige præstationer, som testene leverer, også giver mulighed for, ”at testene virker som pædagogisk værktøj for lærerne” (jf. "De nationale test," Verificeret 2009.03.28).

I maj 2007 tog den daværende ledelse af Dansk Clearinghouse for Uddannelsesforskning initiativ til gennemførelse af et systematisk review med titlen "Pædagogisk brug af test" på baggrund af en Enquete gennemført i november-december 2006.[3] Det er ikke nogen hemmelighed, at inspirationen hertil var introduktionen af nationale test i Danmark. Men det er oplagt at tilsvarende test, som vist ovenfor, også anvendes i Norge og Sverige.

Det skal imidlertid straks understreges, at den foreliggende undersøgelse har et bredere sigte end alene at ligge i forlængelse af de nationale test i de skandinaviske lande. Undersøgelsen bygger, som det vil fremgå nedenfor, på principielt al den primærforskning, der nationalt og internationalt er produceret om de reviewspørgsmål, som formuleres i afsnit 1.4. Undersøgelsen sigter derfor mod at finde evidens for, om og da hvorledes testdata, der stammer fra test af samme type som de nationale test i Skandinavien, kan fungere som et pædagogisk værktøj for lærere.

Det bør derfor straks præciseres, at *den foreliggende undersøgelse ikke direkte behandler de nationale test i Danmark, Norge og Sverige, men er en meta-undersøgelse af internationale primærstudier, der behandler de reviewspørgsmål, som er omtalt i afsnit 1.4.*

Spørgsmålet om "Pædagogisk brug af test" er både meget bredt og ikke særlig klart defineret. I dette afsnit skal dette problemfelt derfor bestemmes nærmere. De tre centrale udtryk – pædagogisk, brug og test - kræver nemlig selvstændige præciseringer og afgræsninger i retning af, hvor bred en tilgang undersøgelsen skal have.

Det skal for det første fastlægges, om fx testdata skal tolkes mere snævert som resultater fra prøver i fx folkeskolefag eller om en bredere tolkning i retning af også at inddrage intelligenstest, holdningstest, sociale test, psykologiske test, osv., skal tillades. Videre skal det fastlægges, om "pædagogisk brug" alene vedrører de anvendelser *lærere* gør, eller om også andre pædagogiske aktører som fx skolepsykologer, lærebogsforfattere, osv. skal inddrages? Endelig skal det besluttes, om den "pædagogiske brug" også omfatter, at resultaterne kan give anledning til politiske tiltag, altså indgår i policyprocessen, med fx ændring af curriculum, fagplaner, skema, nye prøveformer, osv. til følge.

Disse afgræsninger vil nu blive behandlet hver for sig.

1.2.1 Aktører

Der er, som det allerede er nævnt, flere aktører på spil, når man taler om at anvende testdata pædagogisk.

[3] I enqueten indgik 35 personer med virke i folkeskolen, gymnasiet, erhvervsskolen, fagforeninger, embedsmænd i UVM, herunder fagkonsulenter, samt folketingspolitikere. Interviewene blev gennemført af Clearinghouse ud fra en spørgeguide.

Aktørerne er i denne sammenhæng primær tre.[4] Det er primært læreren, der får oplysning om de individuelle testdata i den underviste klasse. Desuden oplyses skolelederen om klassens niveau. Den tredje aktør kan siges at være elevens forældre. I den administrative kæde fra kommune til ministerium, er det ministeriets/ministerens indflydelse, der er den afgørende. Der kan derfor argumenteres for, at de aktører, der kommer på tale i forbindelse med "pædagogisk brug" primært er læreren og "policy-makere". De kommer på banen ved to forskellige typer pædagogisk brug.

Lærernes pædagogiske brug skal først og fremmest ses i to kontekster: Læreren kan anvende de oplysninger, som fremkommer i testen, som redskaber til at vejlede eleven. Vi kunne tale om *"individcentreret pædagogisk brug"*[5]. Denne brug kan være af varieret karakter. Læreren kan beslutte sig til at henvise en elev til særlig behandling, fx hos læsepædagog. Læreren kan beslutte at føre særlig tilsyn med eleven i en længere periode. Læreren kan beslutte at give eleven særlige opgaver – fx mere eller mindre udfordrende end de opgaver, klassen i almindelighed får. Læreren kan beslutte at kontakte forældrene, osv.

Læreren kan også anvende den indsigt, som testdata giver om klassens samlede profil og pædagogiske udvikling som et redskab til at foretage *"klassecentreret pædagogisk brug"*. Denne sidste type anvendelse kan være af mange arter, fx ændring af det faglige niveau, klassedifferentieret undervisning, iværksættelse af temadage, revision af den daglige undervisnings didaktiske udformning, osv.

Policy-makernes (PM) pædagogiske brug kan også ses i to kontekster: PMs reaktioner kan være en følge de politiske omstændigheder, som testdata afstedkommer. Vi kunne tale om *"policy-centreret pædagogisk brug"*. Det er velkendt i Danmark, at fx en undersøgelse af danske elevers læsefærdigheder medførte en panikagtig politisk reaktion og et større forsøgsprojekt om en time om ugen mere undervisning i første sprog i 4.-klasserne (Mejding, 1994). Her var der tale om tiltag, som det var muligt – også økonomisk – at foretage en evaluering af.

Men mange gange vil disse "pædagogiske anvendelser" resultere i fx forslag til ændringer i skoleformer, læseplaner, læreruddannelsesplaner, som ikke kan – eller ikke bliver – ledsages/t af evalueringer med henblik på, om de ønskede virkninger realiseres.

PMs pædagogiske brug kan også bestå i, at man indser nødvendigheden af at udføre et udrednings-, udviklings- eller forsøgsarbejde med assistance fra den pædagogiske forskning. Vi kunne tale om en *"forskningscentreret pædagogisk brug"*. I denne sammenhæng skulle en "pædagogisk brug" af testdatas indsigt gerne føre til, at de tiltag, der anbefales fra forskerside, er evidens- eller forskningsbaserede.

---

[4] Det er en selvfølge, at også eleverne er aktører. Når de ikke er taget med her, er begrundelsen, at de ikke opfattes som direkte ansvarlige for de mulige pædagogiske anvendelser.

[5] Det er veloverlagt, at formuleringen "elevcentreret pædagogisk brug" ikke anvendes. I pædagogisk tradition er "elevcentrering" bl.a. betegnelsen for en velkendt pædagogisk retning på linje med "moderne" og "progressiv" pædagogik. Det centrale i dette review er imidlertid, om interessen primært samler sig om den individuelle elev eller om klassen som kollektiv.

Det systematiske review, som her er gennemført, samler sig om *lærerens pædagogiske brug af testdata,* hvor begge de to ovenfor nævnte perspektiver inddrages, nemlig *individcentreret pædagogisk brug* og *klassecentreret pædagogisk brug*.

Derimod indgår det ikke i denne undersøgelse at redegøre for den policy-centrerede eller forskningscentrerede pædagogiske brug af test eller testdata.

1.2.2 Test

Der er flere mulige måder at afgrænse "test" og "testdata" på. Helena Korp (2003) opregner i bogen *Kunskapsbedömning,* en række af de vigtigste sammenhænge test og testdata kan ses i:

> A. *Det niveau, som resultaterne sammenfattes på*. Hvem vedrører testen? Skal resultatet angive kvaliteten af fx den enkelte elevs præstation, eller af en gruppe af elever, eller skal den gælde for en hel skole, eller for en nation?
>
> B. *Testens betydning*. Skal testens resultat bruges som grundlag for beslutninger af stor betydning enten for det enkelte individ eller for fx en skole eller et uddannelsessystem, eller indgår den i en mindre kritisk sammenhæng? Hvor vigtig er det, at man klarer sig lige godt?
>
> C. *Testens formål*. Gennemføres testen med det sigte at give en karakter eller et bevis, dvs. at formidle resultatet af en vurdering af det, som studeres (fx en elevs viden om et vist fag), eller er formålet med testen i første omgang at bidrage til udvikling og forbedring?
>
> D. *Testens objekt*. Hvad fokuserer testen på? – fx faktaviden, problemløsningsadfærd, æstetisk kompetence, erhvervsegnethed? Eller drejer det sig om samarbejdsevne, om resultatet af testen eller arbejdsprocessen?
>
> E. *Testmetode*. Hvilken type instrument anvender fx en lærer for at finde ud af, hvad en elev kan, ved eller har lært? – fx uformelle observationer, skriftlige prøver, eller praktiske opgaver?
>
> F. *Testprincip*. Hvad sættes vurderingen i relation til? Skal værdien af fx en persons præstation bestemmes i forhold til, hvorledes de øvrige i gruppen præstere, eller i forhold til på forhånd opstillede kriterier? Eller skal testen foretages mere forudsætningsløst og resultere i en kvalitativ beskrivelse?
>
> G. *Hvem udformer og gennemfører testen. Er prøven udformet centralt eller lokalt, er den konstrueret af en undervisende lærer, en lærebogsforfatter eller ved statslig foranstaltning? Og hvis testen vedrører fx en skole eller et uddannelsessystem – er der tale om en ekstern eller intern bedømmelse? Er det den undervisende lærer, som bedømmer sine elever, eller bedømmer eleverne sig selv (selvbedømmelse)? Bedømmer eleverne hinanden (kammeratbedømmelse)?*
>
> H. *Beslutningstagere:* Hvem skal anvende resultatet af testen? – fx undervisende lærere, elever, skole- eller uddannelsesledere, eller såkaldte "policy-makers". (Korp, 2003, 70-71)

Som det vil fremgå, er nogle af disse kategorier allerede omtalt ovenfor, idet afgræsningen til *lærerens* pædagogiske brug svarer til A-kategorien.

Desuden er B-kategoriseringen blev afgrænset til "pædagogisk brug" og det er blevet præciseret, at der her tænkes på den individ- og klassecentrerede brug.

C-kategoriseringen er ikke entydig. Det er muligt at tænke sig et flerfald af formål med at teste. Korp beskriver den grundlæggende forskel ved hjælp af Scrivens velkendte kategorier om summativ og formativ evaluering.

Det er ikke ganske sikkert, om man uden videre kan sige, at de summative evalueringer svarer til de data, som policy-makere anvender, mens omvendt de formative evalueringer direkte er knyttet til lærernes individ- og klassecentrerede pædagogiske brug. Det er i denne sammenhæng værd at understrege, at det ikke er selve testmetoden (jf. E-kategorien), men den anvendelse, som gøres af testen, som er i centrum. Flere forskellige datatyper, udsprunget af forskellige testmetoder, kan bruges som elementer i en formativ evaluering. Det er derfor i sig selv et selvstændigt problem, om kun nogle bestemte testmetoder – og da hvilke – er anvendelige som baggrund for formative evalueringer. Det er ligeledes et åbent spørgsmål, hvorledes resultatet fra en formativ evaluering hænger sammen med bestemte didaktiske tiltag (pædagogisk brug).

Andre mulige formål med test, som fx diagnostiske test og prædiktive/proskriptive test, kan være relevante. I begge disse to eksempler indgår en empirisk tese om, at testmetodens data tillader en tolkning om henholdsvis elevens indlæringsmæssige tilstand eller fremtidige pædagogiske udvikling. Begge disse aspekter kan indgå i læreres pædagogiske brug af testdata, idet det diagnostiske peger mod bestemte mulige udgangspunkter for didaktiske tiltag, som samtidig hævdes vil medføre pædagogiske forbedringer, det proskriptive. For begge disse to præciseringer gælder det imidlertid, at de tvangfrit kan indordnes som præciseringer knyttet til forestillinger om, på hvilken måde formative evalueringer kan føre til pædagogisk brug.

En beslutning om D-kategorien er central. Tager man udgangspunkt i de data, som de nationale test opsamler, ser det ud til, at de i langt de fleste tilfælde samler sig om faktaviden og færdigheden i at løse faglige problemer. Tilføjer man, hvilke fag der faktisk testes i, kan også dette præciseres til testmetoder, der tester faktaviden og problemløsningsadfærd i bestemte fag, og som giver argumenter for, at elevens pædagogiske præstationer kan forbedres ved at iværksætte bestemte didaktiske tiltag.

E-kategorien kan også præcisionsmæssigt styrkes, hvis man tager udgangspunkt i, at de primære testdata, der foreligger, stammer fra nationale test.

Anvendes samme afgrænsningsprincip som ovenfor, kan F-kategorien bl.a. karakteriseres ud fra, om prøverne anskues ud fra et relativt eller absolut perspektiv, dvs. er henholdsvis norm- eller kriterierelaterede test. Ved nationale test vil begge kategorier normalt kunne anvendes. Man vil både kunne bestemme, hvorledes den enkelte elev er placeret i forhold til klassen, og hvorledes klassen er placeret i forhold til landets øvrige klasser. Man vil også få oplysninger om (jf. fx den danske nationale tests begreb om "præstationsprofiler"), på hvilke områder en given elev ikke lever op til de opstillede kriterier. Set ud fra en individcentreret pædagogisk brug vil de kriterierelaterede data have størst interesse, mens man ud fra en klassecentreret brug måske vil have større interesse i data, der er normrelaterede. Befinder min klasse sig under eller over niveau sammenlignet med landsgennemsnittet? I første tilfælde bør særlige tiltag måske sættes ind.

G-kriteriet er ligeledes, hvis nationale test lægges til grund, fastlagt. Der er tale om centralt besluttede og undertiden - som i det danske tilfælde - også om centralt administrerede test.

Om H-kategorien er allerede sagt tilstrækkeligt i afsnit 1.2.1 om aktører.

På baggrund af den ovenstående overvejelse kan det nu præciseres, at det systematiske review, som her præsenteres, beskæftiger sig med *formative evalueringer*, hvor data samler sig om at bestemme *faktaviden* og *problemløsningsadfærd*.

Endelig søger vi efter testmetoder, som – efter analyse – kan identificeres svarende til dem, der anvendes i *nationale test i Skandinavien*.

1.2.3 Brug

Den afgrænsning, der her er på tale, kan også bestemmes som en afgrænsning af den form for intervention, lærere bruger. Denne afgrænsning er nært knyttet til de overvejelser, som er fremsat i afsnit 1.2.1 om aktører. Her er det allerede foreslået, at det systematiske review i snævrere forstand samler sig om lærerens pædagogiske brug af testdata og at denne brug desuden kan anskues som enten individ- eller klassecentreret.

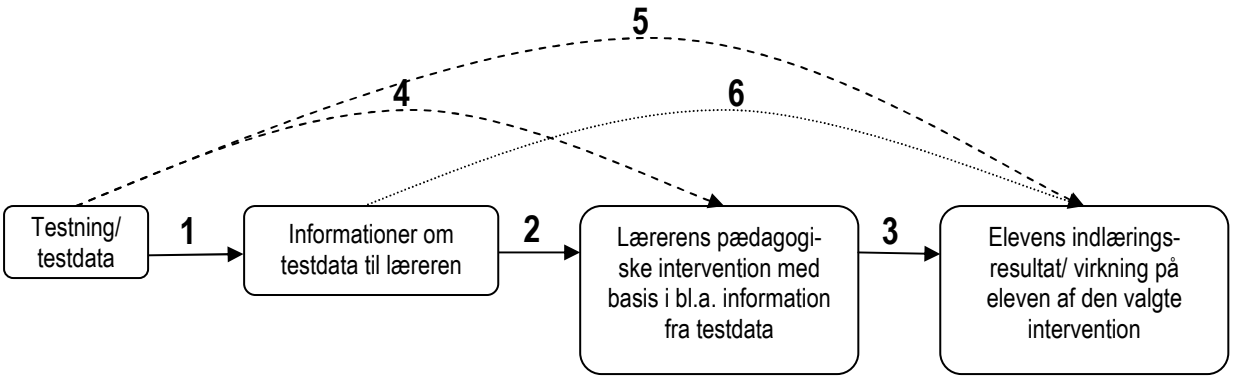
I og med at det systematiske review tager udgangspunkt i fremkomsten af nationale test, hvor bestemte fag er i centrum, jf. Tabel 1.1, Tabel 1.3 og Tabel 1.4, forekommer det ligeledes rimeligt at overveje, om det er muligt at præcisere interventionens karakter ved at koncentrere sig om de didaktiske tiltag, som kendes fra de berørte fags fagdidaktik.

Endelig kan man tale om en mere vag form for "brug" af test. Både i politiske og videnskabelige sammenhænge optræder forestillingen om, at indførelsen af test kan begrundes med forventningerne til de indvirkninger, som læreres og elevers bevidsthed om, at der testes, vil have, altså at de er vidende om, at klassen testes. Forestillingen er, at lærere vil lægge mere arbejde i udformningen af undervisningen, når de ved, der skal testes, og at eleverne tilsvarende vil anstrenge sig fagligt af samme grund.

Vi kan derfor tale om, at man kan forestille sig, at test og testdata kan "bruges" pædagogisk på to måder. I det første tilfælde sker brugen, efter at testen er gennemført og testdata foreligger. I det andet tilfælde sker brugen, før testen er gennemført og testdata endnu ikke foreligger. I sidste tilfælde knytter brugen sig annonceringen af hændelsen 'der skal gennemføres test'. Et heuristisk princip til bestemmelse af disse to former for brug, er at spørge, om brugen ligger før eller efter fremkomsten af testdata.

1.3 Model

Til at organisere forskningen om pædagogisk brug af test i dette systematiske review kan det være hensigtsmæssigt at opstille en konceptuel model for tests og testdatas pædagogiske indflydelse. I denne model koncentreres interessen alene om to aktørgrupper, nemlig lærere og elever. Overordnet set optræder testning og resultatet heraf: testdata, som uafhængige variable, mens den afhængige variabel repræsenteres ved resultatet af disse variablers indflydelse på elevernes læring (pupil achievement).



Figur 1.1: Model af sammenhænge mellem testning/testdata og elevlæring

Vi kan desuden eksplicitere indflydelsen fra testning og testdata til elevers læring som forløbende ad tre veje, nemlig to veje, hvor selve fænomenet 'at testning indføres' virker på henholdsvis læreren og eleven, og en vej, hvor den information, der rummes i testdata, via lærerens pædagogiske/didaktiske tiltag indvirker på elevers læring, jf. afsnit 1.2.3. Figur 2.1 illustrerer modellen og de tre veje, som indflydelsen på elevens læring kan ske ad. Den tager udgangspunkt i hændelsen 'testning af elever'. Dette resulterer i generering af testdata, der tilflyder læreren som en information om de enkelte elevers og hele klassens faglige præstationer. I modellen forudsættes det videre, at informationen om testdata indgår som et blandt flere forhold, der begrunder lærerens didaktiske beslutninger, og at den besluttede intervention har betydning for elevens indlæringsresultat. I modellen er denne "vej" repræsenteret ved de fire tekstbokse, der er forbundet med pilene 1, 2 og 3.

Men forud for fremkomsten af testdata præsenteres elever og lærer som nævnt for en anden, mere uspecifik hændelse, nemlig informationen om, 'at der skal gennemføres en test'. I modellen er virkningen heraf angivet ved pilene 4 og 5, nemlig som en faktor af betydning for lærerens didaktiske beslutninger, der ikke baserer sig på information om testdata, men alene på viden om, at eleverne skal testes, og som en faktor af betydning for elevernes forberedelse af testen.

Endelig er det fra den eksisterende forskning velkendt, at lærerens forventninger til elevernes præstationer i det lange løb virker ind på de enkelte elevers faglige dygtighed, en effekt der med rimelighed kunne betegnes en fjerde vej. Denne forventning kan basere sig på lærerens erfaring med den enkelte elev eller med hele klassen. Men den kan også grunde sig i andre forhold. I et pædagogisk eksperiment, der blev udført af Rosenthal & Jacobson (1977), blev det påvist, at alene lærerens "rene" forventning om en elevs dygtighed alt andet lige i sig selv kunne forklare, at eleven i det lange løb blev dygtig, mens forventningen om en anden elevs manglende dygtighed alt andet lige i det lange løb medførte, at eleven blev mindre dygtig. Dette resultat er blevet bekræftet i senere forskning, (Jussim & Harber, 2005; Rosenthal & Jacobson, 1992).

Fokus i det foreliggende systematiske review er på de relationer, der repræsenteres med modellens pile 1,2, 3, 4 og 5, mens det systematiske review ikke inddrager den relation, der i modellen repræsenteres ved pilen 6.

1.4 Formål

På denne baggrund kan formålene med det systematiske review formuleres således:

> *Hvordan kan grundskolelæreres individ- og klassecentrerede brug af data fra test forbedre læreres didaktiske og/eller fagdidaktiske tiltag i klasser med almindelige elever? – spørgsmålet afgrænses til alene at inddrage testtyper, som indgår i de nationale test i de nordiske lande, og*
>
> *Hvordan indvirker indførelsen af testning på læreres didaktiske beslutninger og elevers læringsadfærd?*

Hermed kan opgavens reviewspørgsmål formuleres på følgende måde:

> *Systematisk forskningskortlægning*: Hvilken empirisk forskning (primærforskning) er gennemført til belysning af grundskolelæreres individ- og klassecentrerede brug af testdatas forbedring af læreres didaktiske og/eller fagdidaktiske tiltag i klasser med almindelige elever? Og
>
> Hvilken empirisk forskning (primærforskning) er gennemført til belysning af virkningerne af introduktion af testning på læreres didaktiske beslutninger og elevers læringsadfærd?
>
> *Systematiske synteser*: Hvilken evidens er der for, at grundskolelæreres individ- og klassecentrerede brug af testdatas forbedrer læreres fagdidaktiske tiltag i klasser med almindelige elever? Og
>
> Hvilken evidens er der for, at introduktion af testning influerer på læreres didaktiske beslutninger og elevers læringsadfærd?

1.5 Reviewgruppe

I forbindelse med løsningen af denne opgave har der været nedsat en reviewgruppe med følgende medlemmer:

> Professor Peter Allerup, Institut for Læring, Aarhus Universitet,
> Professor Hanne Leth Andersen, Center for Undervisningsudvikling, Aarhus Universitet
> Institutleder Jens Dolin, Institut for Naturfagenes Didaktik, Københavns Universitet
> Lektor, fil.dr. Helena Korp, Högskolan Väst
> Postdoc, dr.polit. Rolf Vegar Olsen, Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo

Reviewgruppen har deltaget ved datauddragning/genbeskrivelse af de kortlagte undersøgelser og i samarbejde med Clearinghouse udarbejdet den tekniske rapport. Der har ikke været interessekonflikter for noget reviewgruppemedlem i tilknytning til denne rapport.

2 Metoder anvendt i reviewet

2.1 Design og metode

Dette systematiske review er gennemført på det grundlag, der er beskrevet i Dansk Clearinghouse for Uddannelsesforsknings konceptnotat (http://dpu.dk/Clearinghouse.**).

Det er skabt ved gennemløb af en række trin og processer på eksplicit og transparent vis. Dette vil fremgå af den følgende rapport.

Dansk Clearinghouse for Uddannelsesforskning har valgt i denne sammenhæng at anvende et særligt software til arbejdet, EPPI-Reviewer. Softwaren er nærmere beskrevet på producentens hjemmeside:

http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/EPPI-Reviewer_Feb_06.pdf

Det transparente og eksplicitte er også et princip efter gennemførelsen af det systematiske review. Her er genbeskrivelserne af de undersøgelser, som reviewet er baseret på, gjort tilgængelige i denne database:

http://eppi.ioe.ac.uk/webdatabases/Intro.aspx?ID=6

2.2 Begrebsmæssige afgræsninger

I dette systematiske review er der i udgangspunktet ledt efter svar på et snævert spørgsmål:

Hvordan kan grundskolelæreres individ- og klassecentrerede brug af data fra test forbedre læreres fagdidaktiske tiltag i klasser med almindelige elever?

Da det undervejs i reviewprocessen viste sig, at den snævrere bestemmelse af reviewspørgsmålet kun gav relativt få fund blev det besluttet at supplere med en lidt bredere bestemmelse af spørgsmålet:

Hvilke virkninger forbundet med testning er der på lærere og elever i grundskolen?

Reviewprocessen er forløbet på en måde, der sikrer at undersøgelser, der kan besvare

både det ene og det andet spørgsmål, faktisk findes.

Spørgsmålene afgrænses på følgende vis:

Ved 'test' skal forstås egentlige tests, ikke mere eller mindre uformelle tilgange til at få data om elever. Alene testtyper, som indgår i de nationale test i de nordiske landes grundskoler, kan indgå. Dette betyder, at der alene sigtes mod skolastiske tests. Skolastiske tests findes dog i mange former.

De 'aktører' der er fokus på er elever og lærere. Skoleledere, forældre, offentligheden, politikere etc er ikke aktører i dette systematiske review.

'Grundskole' er bestemt som almindelige skoler (ikke specialskoler) på primært eller sekundært niveau.

'Fagdidaktiske' skal i dette review forstås som omhandlende følgende fag: læsning/sprog, matematik og naturvidenskaber.

Der er ikke på forhånd taget stilling til at nogen særlige forskningsdesign er bedre til at besvare reviewspørgsmålene end andre.

Således er årsag eller baggrund (testning eller information fra tests) begrebsmæssigt afklaret. Dette gælder ligeledes kontekst (primær eller sekundærskoler i læsning/sprog, matematik eller naturvidenskaber) og aktører (lærere og/eller elever). Derimod er virkning ikke på forhånd afklaret. Dette systematiske review handler om at finde frem til/navngive virkninger.

2.3 Søgning

De begrebsmæssige afgrænsninger er det grundlag, hvorpå det er besluttet hvilke databaser og ressourcer, der skal afsøges og med hvilke profiler.

Søgninger er foretaget af Clearinghouse og omfatter perioden 1980-2008. Reviewgruppen har haft lejlighed til at kommentere alle søgninger. Desuden har de haft mulighed for at supplere søgninger med referencer. Gruppen har udnyttet denne mulighed.

I Tabel 2.1 herunder er vist hvilke databaser/ressourcer, der er afsøgt, hvornår de er afsøgt og med hvilke antal fund.

Ressource	Søgedato	Hits
ERIC	070713	3279
BEI	070718	365
AEI	070718	551
CBCA-education	070718	59
Psychinfo	070718	239
Fisbildung	070719	299
JYKDOK	070725	193
Evidensbasen	070726	7
News Alerts DPB	continuous	12
Libris	070726	335
Dansk pædagogisk base	070831	229
NORBOK	070727	20
References from review group	continuous	9

Tabel 2.1: Databaser/ressourcer, der er afsøgt

Disse ressourcer kan kort beskrives således:

ERIC er verdens største pædagogiske database. Den er søgt i versionen hos CSA

BEI, British Education Index, er den britiske database med pædagogisk litteratur. Den har et vist, men ikke fuldstændigt overlap med ERIC.

AEI, Australian Education Index er den australske database med pædagogisk litteratur.

CBCA-education, Canadian Business and Current Affairs- education, er den canadiske pædagogikdatabase. Den dækker både enegelsk og franskproget litteratur.

Psychinfo er den førende internationale psykologiske datababase. Den er afsøgt I versionen hos CSA.

Fisbildung er databasen for den tysksprogede pædagogiske litteratur, hvad enten den kommer fra Tyskland eller andre lande.

Jykdok er bibliotekskataloget hos det finske ansvarsbibliotek for pædagogik. Her findes den finske pædagogiske literatur.

Evidensbasen er Dansk Clearinghouse for Uddannelsesforsknings database, som indeholder beskrivelser af systematiske reviews fra hele verden.

News alert fra DPB er en tjeneste der giver indholdsfortegnelser fra nye pædagogiske tidsskrifter. Disse er undersøgt for relevante nye undersøgelser.

Libris indeholder den svenske bogfortegnelse, som også indeholder svensk pædagogisk litteratur.

Dansk Pædagogisk Base er en database produceret af Danmarks Pædagogiske Bibliotek, som indeholder dansk pædagogisk litteratur.

NORBOK er norsk bogfortegnelse, som også indeholder norsk pædagogisk litteratur.

References from Review group er de referencer som medlemmer af reviewgruppen har suppleret med undervejs.

Ved at anvende disse ressourcer er der i reviewet åbnet for inddragelse af pædagogisk forskning fra den største del af den industrialiserede verden. I sproglig henseende er der søgt efter forskning på engelsk, tysk, fransk og skandinaviske sprog.

Almindeligvis kan et systematisk review, med Clearinghouse's arbejdsform, gennemføres på ca. 1 år. Arbejdsprocessen i dette review er af forskellige grunde blevet længerevarende. Derfor har det været nødvendigt at gentage alle søgninger for at supplere med den nyeste litteratur. Disse ekstra søgninger er anført i Tabel 2.2.

Database/Ressource	Søgedato	Hits
Ericupdate	09/07/2008	199
BElupdate	09/07/2008	19
AElupdate	09/07/2008	10
CBCAupdate	09/07/2008	7
Psychinfoupdate	09/07/2008	70
Fisbildungupdate	09/07/2008	17
JYKDOKupdate	10/07/2008	3
Evidensbasenupdate	09/07/2008	1
Librisupdate	10/07/2008	24
Dansk pædagogisk base update	10/07/2008	38
NORBOKupdate	10/07/2008	1

Tabel 2.2: Ekstra søgninger august 2008

De faktisk anvendte søgeprofiler er udformet, så at reviewets tema er udtrykt i hvert af de afsøgte ressourcers grænseflade (emneord, klassifikationssystemer og/eller fritekstsøgning).

Der er søgt efter empirisk forskning - dog ikke efter bestemte forskningsmetodiske tilgange. Der er endvidere søgt på en måde, hvor test og testning er ekspliciteret og specificeret. Derimod er der stort set ikke i søgeprocessen specificeret virkninger i det pædagogiske rum. Der er søgt på denne vis, dels fordi virkningerne kun i begrænset udstrækning på forhånd kan navngives, dels fordi virkninger vil være med i mængden af hits, som handler om forskellige testformer. Derfor er der i søgningerne fundet overordentlig meget materiale, som senere i screeningen viste sig ikke-relevant.

Testformer, der er søgt efter, er dels en række almene betegnelser for skolastiske testformer, dels testformer i bestemte skolefag. Derudover er der dog søgt efter to navngivne virkninger forbundet med testning. Dette er dels 'backwash' fænomenet, der angår en tests virkning på det (forudgående) undervisningsforløb, dels 'konsekvens validitet', der er beregningen af testvaliditet set i forhold til testanvendelse.

De meget omfattende søgeprofiler fra alle søgninger findes i Kapitel 6, Appendiks 1.

2.4 Screening

Som det fremgår, er der søgt på en måde, der både skulle sikre, at den ønskede litteratur faktisk findes, men også på en måde, der er forbundet med en forventning om, at meget af det fundne vil være ikke-relevant. Screening for relevans af alle fund er derfor nødvendig. Screeningen er udført af medarbejdere på Clearinghouse i samarbejde med Reviewgruppen.

Forud for screening blev der fjernet dubletter. Det er uundgåeligt, at der findes dubletter i en søgeproces af den art der her er gennemført. I alt er der fjernet 270 dubletter.

De resterende 5716 fund er blevet screenet i en proces med to faser: referencescreening og fuldtekstscreening.

2.4.1 Fase 1: referencescreening

Alle referencer i EPPI reviewer er sorteret efter de kategorier som findes i Tabel 2.3.

Der er ekskluderet og inkluderet systematisk hierarkisk. Først er det første kriterium, *wrong test scope*, søgt anvendt på referencen, hvis dette ikke kunne anvendes er det derefter vurderet, om der kan ekskluderes på næste kriterium, *wrong institution*. Etc.

Især de nordiske referencer, men også en del af de øvrige giver for lidt information til sikker beslutning om inklusion/eksklusion sådanne referencer er i processen midlertidigt blevet placeret under *Marker Insufficient information at present*. Her er der undervejs ved nye søgninger ledt efter yderligere informationer om referencen. Der er kun blevet ekskluderet referencer, hvor der har været et sikkert grundlag for at gøre det.

Fase 1 screeningen er i denne review proces forløbet to gange, da der er foretaget søgninger to gange. Den samlede fase 1 screening endte med 119 inkluderede referencer.

Reason for inclusion/Exclusion	Reason described	Number
EXCLUDE Wrong test scope	Not on the effects or use or misuse of test in schools	5084
EXCLUDE wrong institution	Not on activities in primary or secondary school	86
EXCLUDE Wrong paper	Not a paper with data from empirical research: Editorials, commentaries, book reviews, policy documents, resources, guides, manuals, bibliographies, opinion papers, theoretical papers, philosophical papers, research-methodology papers	372
EXCLUDE Wrong research	Not offering data from original research i.e. only summarizing research done by others. (However systematic reviews can be included)	82
EXCLUDE Wrong actors	Not offering information on how teachers or pupils are affected by or acts with tests	17
MARKER Insufficient information at present	The document description is not sufficient to warrant inclusion/exclusion.	0
MARKER Overview	Excluded but a document which provides historic or systematic overview of the review theme.	74
INCLUDE NOT in 1-8	Transfer of references to review	72

Tabel 2.3: Referencer i EPPI reviewer sorteret efter kategorier

Nærmere iagttagelse af eksklusionskriterierne viser, at der i dette systematiske ikke er ekskluderet på grundlag af forskningsdesignmæssige overvejelser. Dette er i god overensstemmelse med en bredere tolkning af reviewspørgsmålet, hvor mange forskellige forskningsdesign principielt kan komme på tale. Der er dog krav om at der skal være tale

om en undersøgelse. F.eks. er forsøgs- udviklingsarbejder ikke ekskluderet, for så vidt de er rapporteret som undersøgelse.

Ligeledes skal det understreges, at vurdering af forskningskvalitet ikke er en del af screeningen. Dette moment indgår først i forlængelse af genbeskrivelse af undersøgelserne. Se herom senere.

2.4.2 Fase 2: fuldtekstscreening

Alle referencer inkluderet i fase 1 skaffes derefter, hvad enten det handler om en artikel, bog eller rapport. Fase 2 screeningen foretages i forhold til den fulde tekst af disse.

For yderligere at belyse screeningsprocessen, vil de enkelte eksklusionskriterier, i.e. det grundlag undersøgelser er fjernet på, nøjere blive omtalt her.

'Wrong test scope' er først og fremmest langt de fleste referencer, der handler om konstruktion af test- ikke om hvad der sker, når tests anvendes. Desuden indgår heri undersøgelser, hvor tests alene sammenlignes med andre test. Desuden omhandler dette undersøgelser, hvor det ikke er egentlige tests, der undersøges, men mere uformelle målemetoder, som eksempelvis lærerkonstruerede spørgerammer. Undersøgelser, hvor testen alene spiller en rolle for analyse af et andet pædagogisk forhold, f.eks. en undervisningsmetode er også ekskluderet på dette kriterium.

'Wrong institution' kriteriet handler om, at de indgående institutioner ikke er almindelige primær eller sekundærskoler. Specialskoler og uddannelsesinstitutioner på andet niveau er fjernet på dette kriterium.

'Wrong Paper'. Dette angår forholdet, at den fundne reference ikke rapporterer forskning, Der kan f.eks. være tale om lærebøger, rene forskningsmetodiske eller videnskabsteoretiske diskussioner eller rent diskuterende fremstillinger.

''Wrong research' kriteriet er anvendt, når dokumentet ikke er en redegørelse for forfatterens egen forskning. Systematiske reviews er dog inkluderet.

'Wrong actors' er anvendt til at ekskludere dokumenter, hvor virkninger af test eller testinformation ikke er undersøgt på elever eller lærere.

Den samlede screeningsproces er vist i Figur 2.1.

2.5 Genbeskrivelse/datauddragning af studier

De 71 dokumenter, der kunne skaffes angår 61 undersøgelser. Det er disse 61 undersøgelser, der er genbeskrevet og vurderet. Dette arbejde er sket i et samarbejde mellem Reviewgruppen og medarbejdere fra Clearinghouse.

Genbeskrivelse og vurdering er sket ved at anvende 'EPPI Centre data extraction and coding tool for education studies v2.0'. Dette system er udviklet af EPPI centre ved Institute of Education, Londom University. Et eksempel på anvendelsen af genbeskrivelsessystemet er givet i Kapitel 7: Appendix 2.

Ved at anvende det samme system på alle undersøgelser, kommer dette til at fungere som et tertio comparationis. Dvs. det bliver muligt at sammenligne og sammenholde data og resultater fra forskellige undersøgelser.

Genbeskrivelsessystemet er opbygget som en ramme af spørgsmål, som skal besvares for hvert enkelt studie. Systemet er strukturelt opbygget i sektioner, som indeholder spørgsmål, som er ordnet i multiple choice svarmuligheder. Overalt er der mulighed for at koble uddybende noter til givne afkrydsningssvar. Systemet dækker en undersøgelses formål, kontekst, design, metode, resultater og forsknings- og rapporteringskvalitet.

Genbeskrivelserne af den enkelte undersøgelse er gennemført i et samarbejde mellem en medarbejder fra Clearinghouse og et medlem af reviewgruppen. I processen har der således altid været mindst to, der har set på/arbejdet med genbeskrivelsen. Processen er ligeledes gennemført med transparens for alle i Clearinghouse og Reviewgruppe. På denne vis er genbeskrivelser og vurderinger kvalitetssikret.

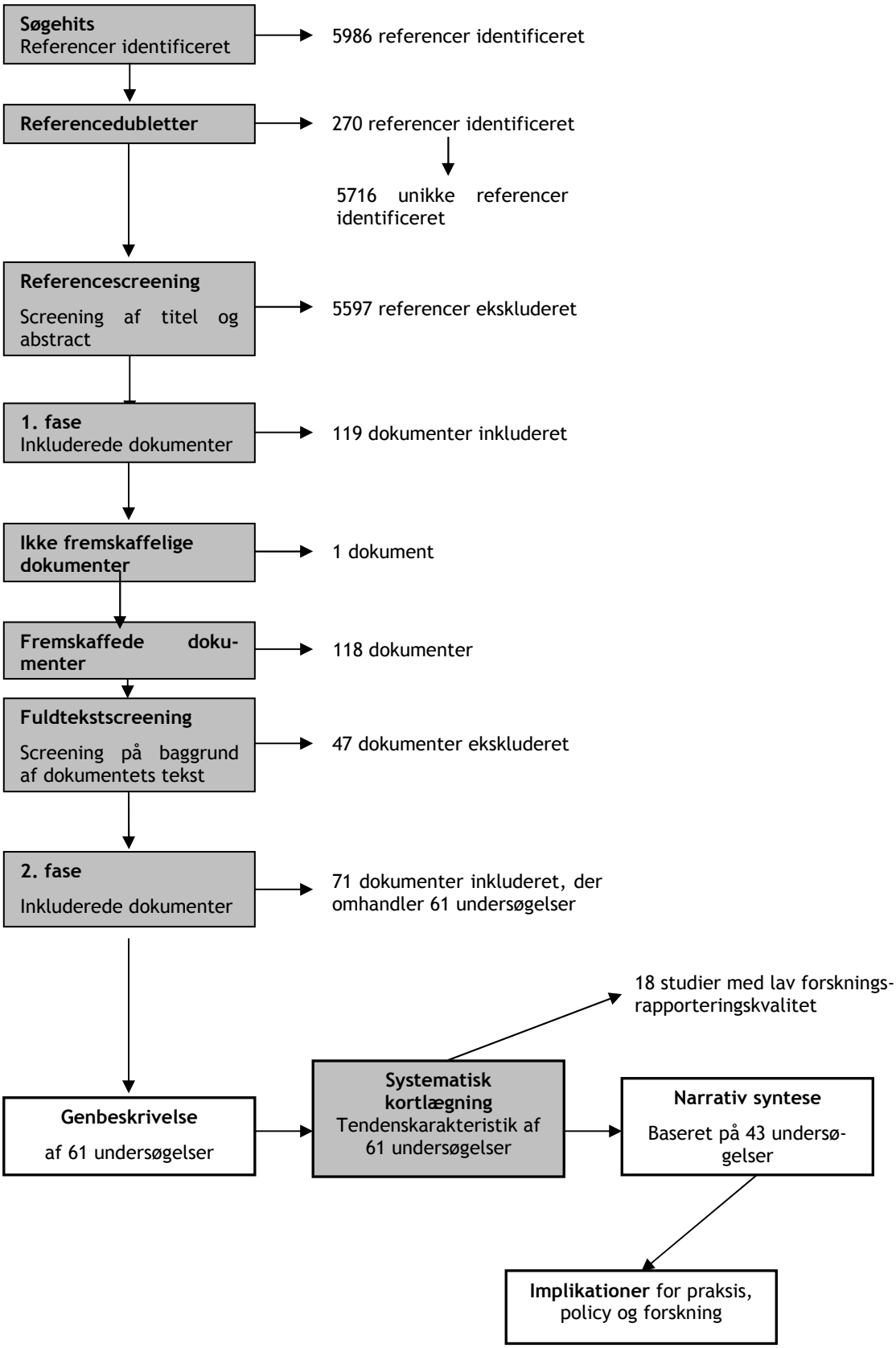
EPPI systemet er velegnet til analyse af forskning af den art, der indgår i dette review, hvor langt den største del er kvantitativt orienteret. EPPI systemet kan dog også håndtere kvalitativt orienteret forskning.

For at ramme de særlige momenter i det reviewspørgsmål, der er i fokus her, er EPPI systemet suppleret med et reviwspecifikt genbeskrivelsessystem. Dette er opbygget på samme måde som EPPI systemet og dækker: test, testning og testkarakteristika samt retning af testens virkning/påvirkning. Et eksempel på anvendelse af det reviewspecifikke genbeskrivelsessystem findes i Kapitel 7, Appendiks 2.

Det er med baggrund i disse 61 genbeskrivelser, at der kan gives en redegørelse for forskningen på feltet.

2.6 Samlet oversigt over reviewprocessen

Figuren på næste side viser logistikken i dette systematiske review fra søgninger til syntese og implikationer af syntesen. Figuren viser også hvilke delprocesser Clearinghouse primært har varetaget,, de grå bokse, og hvilke delprocesser Reviewgruppe og Clearinghouse har været fælles om, de hvide bokse.



Figur 2.1: Filtrering af referencer fra søgning over kortlægning til syntese

3 Forskningskortlægning og forskningsvurdering

Dette kapitel gennemgår karakteristika ved den fundne forskning, som kan give svar på reviewspørgsmålene, der findes på side 33. Først behandles almene karakteristika som publiceringssprog og undersøgelsesland. Dernæst ses der nærmere på studiernes metodiske og designmæssige karakter. Så behandles studierne i forhold til reviewspørgsmålenes momenter:

Test(type), fag og virkninger samt de to reviewspørgsmål. Afslutningsvis redegøres der for kvaliteten af de 61 undersøgelser, der indgår i analysen.

Forskningskortlægningen og -vurderingen er baseret på den omfattende genbeskrivelse og kvalitetsvurdering, som er i Kapitel 2.

3.1 Almen karakteristik

Der er søgt på en måde, der muliggør fund af studier gennemført i overordentlig mange lande og publiceret på overordentlig mange sprog.

Fordeling af studer på de lande, hvor de er gennemført, er vist i Tabel 3.1.

Undersøgelsesland	Antal studier
USA	32
Kina/Hong Kong/Taiwan	6
U.K.	5
Israel	4
Australien	3
Canada	2
Sverige	2
Danmark	1
Svejts	1
Trinidad Tobago	1
Sri Lanka	1
Japan	1
Uspecificerede eller mange lande	4

Tabel 3.1: Studierne fordelt på de lande, de er gennemført i (N=61 studier)

Som det fremgår, er ca. halvdelen af studierne gennemført i USA. De resterende fordeler sig på mange forskellige lande. Antallet af lande er større end 61 fordi enkelte af

studierne er komparative undersøgelser, hvori indgår flere lande. Gruppen uspecificerede eller mange lande indeholder blandt andet de systematiske reviews, der indgår i denne undersøgelse.

Medens der er en vis spredning i studierne på undersøgelsesland, så er dette ikke tilfældet på publiceringssproget. Som det ses af Tabel 3.2, er næsten alle studier publiceret på engelsk.

Publiceringssprog	Antal studier
Engelsk	58
Tysk	1
Dansk	1
Svensk	1

Tabel 3.2: Studiernes publiceringssprog (N=61 studier)

3.1.1 Metodisk/designmæssige karakteristik

Der er gennemført søgning og screening på en måde som ikke udelukker nogle designs frem for andre. De 61 studiers design fordeler sig som i Tabel 3.3.

Forskningsdesign	Antal studier
Case study	9
Cohort study	8
Cross-sectional study	16
Document study	6
Ethnography	6
Experiment with non-random allocation to groups	5
Methodological study	1
One group post-test only	3
One group pre-post test	4
Other review (non systematic)	1
Secondary data analysis	2
Systematic review	3
Views study	26

Tabel 3.3: Forskningsdesign anvendt (N=61 studier)

Adskillige studier betjener sig af flere designs. Derfor er antallet af designs (90) større end antallet af studier (61).

Designmæssigt er der stor spredning i studierne. Views study, der angår aktørers holdninger eller indstillinger til fænomener er et moment, der indgår i 43 % af studierne. Cross-sectional study, hvor aktører med forskellige aldre undersøges på én gang, anvendes i 25 % af studierne. Case study, hvor et enkelt fænomen undersøges kvalitativt i dybden indgår i 15 % af studierne.

Langt de fleste studier har en kvantitativ tilgang til feltet. Under en fjerdedel af studierne har et kvalitativt moment i sit design (case study, ethnography).

Ses der nøjere på studierne med eksperimentelt eller quasi-eksperimentelt design, viser det sig, at der ikke er gennemført randomiserede kontrollerede undersøgelser. Da begge reviewspørgsmål omhandler effekt, hvor et sådant design må anses for velegnet, er dette ikke optimalt.

3.2 Reviewspecifik karakteristik

Her vil de forskellige momenter i reviewspørgsmålene blive behandlet. Først ses der på karakteristika ved de anvendte tests. Dernæst ses der på virkningskategorier i studierne.

En nøjere analyse af de fundne studiers indhold kunne imidlertid suppleres med en karakteristik af de mange studier, som i reviewprocessen er blevet ekskluderet. Her forholder det sig sådan, at langt den overvejende del af forskningen er orienteret mod, at

skabe pålidelige tests. Det der sker eller kan ske efter introduktion af et givet testsystem i skolen, er kun i meget beskedent omfang undersøgt.

3.3 Karakteristika ved de anvendte test

I forbindelse med dette review er der udformet en række ganske detaljerede spørgsmål til brug for en karakteristik af de undersøgte tests.

Studiernes fordeling på overordnet testformål fremgår af Tabel 3.4.

Da mange af de anvendte test har flere formål forbundet med testningen er der flere testformål (139) end studier (61). Først og fremmest kan det her konstateres, at materialet er heterogent. Sigtet med testningen fordeler sig på flere formål. 66 % af studierne er rettet mod præstationsmåling i et fag eller fagområde. 34 % af studierne har et diagnostisk sigte med testningen. Der er 31 % af studierne, hvor testningen er rette mod kvalitetssikring og accountability. 31 % af studierne anvender tests, der sigter mod at informere den pædagogiske praksis. 26 % af studierne har tests, der sigter mod at give elever karakterer eller mod at sammenligne elevers præstationsniveau med andre elevers. Støtte til elevernes udvikling af indlæringsstrategier er sigtet for testanvendelsen i 16 % af studierne. Gruppen 'andet', hvor 21 % af studierne er repræsenteret, fordeler sig på flere eksplicitte formål som streaming af elever, adgangseksamen, ændring af pædagogisk praksis som led i en uddannelsesreform eller testning som et bevidst redskab for forandring af pædagogisk praksis.

Overordnet testformål	Antal studier
Accountability/måling af pædagogisk kvalitet	19
Diagnose af elever	21
Ligestille karakterniveauer i og mellem skoler	1
Give elever karakterer eller sammenholde elevers præstationsniveau med andre	16
Informere den undervisningsmæssige praksis	19
Måling af en elevs præstationsniveau i et fag eller fagområde	40
Støtte elevers udvikling af læringsstrategier	10
Andet	13

Tabel 3.4: Studiernes fordeling på overordnet testformål (N=61 studier)

Da mange af de anvendte test har flere formål forbundet med testningen er der flere testformål (139) end studier (61). Først og fremmest kan det her konstateres, at materialet er heterogent. Sigtet med testningen fordeler sig på flere formål. 66 % af studierne er rettet mod præstationsmåling i et fag eller fagområde. 34 % af studierne har et diagnostisk sigte med testningen. Der er 31 % af studierne, hvor testningen er rette mod kvalitetssikring og accountability. 31 % af studierne anvender tests, der sigter mod at informere den pædagogiske praksis. 26 % af studierne har tests, der sigter mod at give

elever karakterer eller mod at sammenligne elevers præstationsniveau med andre elevers. Støtte til elevernes udvikling af indlæringsstrategier er sigtet for testanvendelsen i 16 % af studierne. Gruppen 'andet', hvor 21 % af studierne er repræsenteret, fordeler sig på flere eksplicitte formål som streaming af elever, adgangseksamen, ændring af pædagogisk praksis som led i en uddannelsesreform eller testning som et bevidst redskab for forandring af pædagogisk praksis.

De kompetencer, som de undersøgte tests måler, er anført i tabellen herunder. Der er flere studier, hvor testene måler forskellige kompetencer. Derfor er antallet af registrerede kompetencetyper (118) større end antallet af undersøgelser (61).

Tabel 3.5 viser stor spredning i de kompetencer som testene undersøger. Mestring af specifikt indhold behandles i tests, der indgår i 61 % af studierne. Procedurale færdigheder indgår i testene i 33 % af studierne. 28 % af studierne har tests, hvori indgår problemløsningsfærdigheder. Færdighed i ræsonnement er med i 23 % af studiernes testning. I gruppen andre indgår studier, hvor det kompetencemæssige moment i den anvendte test ikke er tilstrækkeligt klart beskrevet og studier, der har karakter af systematiske reviews, hvor det kompetencemæssige moment er meget bredt.

[*** Kommentar: Gruppen 'Andre' (med oprindeligt 29) er opløst ved at køre en 'summary report (a)' med filtret 'gruppen der har svaret andre på spørgsmålet': 14 af studierne burde have været kategoriseret som mestring af specifikt indhold. Disse er tilføjet dette ovenfor. (der er i alle 14 tilfælde tale om sprog- eller læsefærdighed MSL. Det kunne overvejes at åbne og rette disse 14 studier på dette punkt).]

Testens indhold angår	Antal studier
Æstetiske færdigheder	2
Kreativitet	5
Færdighed i kritisk tænkning	3
Dømmekraft	2
Mestring af specifikt indhold (facts, begreber, teorier)	37
Meta-kognitive færdigheder	3
Færdighed i ræsonnement	14
Problemløsningsfærdigheder	17
Procedurale færdigheder	20
Andre	15

Tabel 3.5: Hvilket indhold angår testene?
(N=61 studier)

Fordelingen af de opgavetyper, som studiernes test betjener sig af, fremgår af Tabel 3.6:

Testens spørgsmål	Antal studier
Lukkede	21
Ikke anvendelig	30
Åbne	18

Tabel 3.6: De opgavetyper, som studiernes test anvender (N=61 studier)

Der er 69 markerede svar på de 61 studier. 49 % af studierne giver utilstrækkelige informationer til, at dette spørgsmål kan besvares sikkert eller de handler om flere forskellige tests, der undersøges (f.eks. i et systematisk review). 34 % af testene anvender lukkede spørgsmål og 30 % gør brug af åbne spørgsmål.

Testenes svarformat fremgår af Tabel 3.7.

Testens svarformat	Antal studier
Mundtligt (udvidet)	6
Aktivitet (performance)	4
Skriftligt (kortfattet)	31
Skriftligt (udvidet)	22
Andet	24

Tabel 3.7: Testenes svarformat (N=61 studier)

Der er i alt 87 svar fra de 61 studier. 51 % af studierne angår tests, som anvender den skriftlig kortfattede form som svarformat. 39 % anvender skriftligheden i længere format. Mundtlighed og aktivitet som svarformat er kun med i få af studierne. Gruppen 'Andet' indeholder 17 studier, hvor der ikke er tilstrækkelig information om dette, samt 8 studier hvor svarformat formen er forskellige blandinger af de enkelte svarformatformer.

Hvem der scorer testen	Antal studier
Kommercielle institutter	1
Lokale lærere	24
Specialuddannede lærere på centralt niveau	20
Andre	26

Tabel 3.8: Hvem scorer testen?
(N=61 studier)

Der er 71 svar fra de 61 studier hvilket markerer at flere af studierne angår test hvor scoring udføres af flere parter. I 39 % af studierne scores der af lokale lærere. I 33 % foretages scoring af centralt placerede specialuddannede lærere. I kategorien 'Andre' er der 24 studier med utilstrækkelige informationer om hvem, der scorer testen. De to sidste af disse nævner: de lokale uddannelsesmyndigheder og universiteterne, der har skabt testen.

Hvem gør brug af eller hvem er testdata rettet mod?	Antal studier
Regeringen/centraladministrationen	11
Skoleleder/skolebestyrelse	17
Lokale skolemyndigheder	23
Lokale lærere	42
Elever og forældre	30
Andre	15

Tabel 3.9: Hvem bruger eller retter testene sig mod?
(N=61 studier)

Det er almindeligt at flere parter gør brug af testdata. Der er 138 svar på de 61 studier. Oftest er testdata rettet mod lokale lærere, 69 % af studierne, eller mod elever og forældre, 49 % af studierne. I det hele taget er langt de fleste af studierne om tests, hvor testdata er rettet mod det lokale. Kun 18 % af studierne er om tests, hvor testene er rettet mod regeringen eller centraladministrationen. Gruppen 'Andre' indeholder 8 studier med utilstrækkelige informationer om denne sag og 7 studier der nævner den institution, der skal modtage eleven som oftest universitetet.

Hvad er på spil?	Antal studier
For eleverne	57
For andre	25

Tabel 3.10: Hvem har testene konsekvenser for?
(N=61 studier)

Spørgsmålet 'Hvad er på spil?' viser hen til begrebet "at stake" på engelsk. Det, der er fokus på her, er konsekvenser, det være sig oplevede eller reelle af at klare sig mere eller mindre godt i en test. Der er 'noget på spil' både for elever *og* andre i 21 af studierne, da der er 82 svar på 61 studier.

Hvad der er på spil fremgår af svarene.

Svaret 'For andre' end eleverne, der gives i 41 % af studierne, rummer for det første 7 studier, hvor denne information er uklar eller uoplyst. De resterende af disse studier nævner lokale lærere, - skoleledere og - skolemyndigheder. Hvor meget, der er på spil, varierer. Nogle studier undersøger tests, der er forbundet med, at der er meget at vinde eller tabe (high stakes). I andre tilfælde er der ikke så åbenbart meget at vinde eller tabe (low stakes). Det må bemærkes, at det, der her er undersøgt, er de officielle muligheder for sanktion og belønning forbundet med testningen. De oplevede muligheder for sanktion og belønning behøver ikke at være i overensstemmelse hermed. Det vil der blive set nøjere på i afsnittet om syntese af studierne. Konkret kan 'high stakes' her dreje sig om økonomiske ressourcer, i form af belønning eller sanktion. Et andet eksempel er, at akkrediteringen kan fratages en skole med for dårlige testdata. For lærerne er der i de tilfælde, hvor resultater offentliggøres på klasseniveau, mulighed for high stakes. Tilsvarende for skoleledere, når resultater offentliggøres på skoleniveau.

Svaret 'for eleverne' er forbundet med testbrugen i 93 % af alle studier. 23 af disse studier giver dog ingen eller uklar information om, hvad der er på spil for eleverne. De resterende 34 giver eksempler både på objektive high stakes tests (eksempel: optagelse eller ej på universitet) og objektive low stakes test (eks: tests der sigter mod planlægning af undervisning - ikke mod at vurdere eleverne).

Også her er det vigtigt at være opmærksom på at elevernes oplevelse af om testen er high stakes eller low stakes ikke behøver at være i overensstemmelse med testens officielle sigte.

Det er også undersøgt, om de test der indgår i studierne, gør brug af en skala og i bekræftende fald af hvilken type. Dette fremgår af Tabel 3.11.

Anvendelse af skala	Antal studier
Ingen skala i brug eller ingen omtale deraf	29
Normbaserede standarder	9
Kvalitative standarder	10
Kvantitative standarder	19

Tabel 3.11: Gør testene brug af en skala?
(N=61 studier)

Der er 67 svar på de 61 studier, så enkelte af studierne handler om tests, hvor forskellige skalatyper tages i brug. 48 % af studierne anvender ingen skala eller omtaler ikke nærmere hvilken skalatype der er i brug. Kvantitative standarder findes i 31 % af studierne, medens kvalitative standarder og normbaserede standarder indgår i henholdsvis 16 og 15 % af studierne.

Det format som testresultatet kommunikeres til eleverne i er anført i Tabel 3.12:

Format for kommunikation af testresultat	Antal studier
Kvantitativ eller standardiseret rapport	23
Kvalitativ eller ikke-standardiseret rapport	10
Ikke anvendelig	32

Tabel 3.12: Testdatas format
(N=61 studier)

Da der er 65 svar på de 61 studier betyder det, at 4 studier både giver en kvalitativ og en kvantitativ tilbagemelding om testdata. I 51 % af studierne kan dette spørgsmål imidlertid ikke besvares. Der er slet ikke eller utilstrækkelige oplysninger om dette i disse. De systematiske reviews, der indgår, omfatter så mange forskellige tests, at det heller ikke er et spørgsmål der kan besvares meningsfuldt.

38 % af studierne indeholder tests, hvor testresultatet kan kommunikeres som en kvantitativ eller standardiseret rapport. 16 % af studierne omhandler test, hvor testdata kan kommunikeres ikke-standardiseret.

3.4 Fag og virkninger

Som det er fremgået af afsnittet om søgning er der søgt efter undersøgelser med virkninger i bestemte fag og fagområder. Dette udelukker ikke, at der kan findes studier, som også rapporterer virkninger i andre fag. De fag, som studierne angår, fremgår af Tabel 3.13.

Fagområde	Antal studier
Kunst	2
Geografi	1
Historie	1
ICT	1
Læse- og skrivefærdighed – førstesprog	17
Læse- og skrivefærdighed – andet og tredjesprog	11
Litteratur	1
Matematik	22
Naturvidenskab	7
Andre fag	7
Ikke anvendelig	3

Tabel 3.13: De fag, som studierne angår
(N=61 studier)

Der er 73 svar på de 61 studier, så en del af studierne angår flere fagområder på én gang.

De sproglige fag, omfattende læse- og skrivefærdighed, både første, andet og tredje sprog samt litteratur, har den største mængde studier med 48 % af alle svar. Matematik med 36 % og naturvidenskab med 11 % kommer derefter. I gruppen 'Andre fag' er der et par studier, som inddrager skolefaget samfundsfag. De resterende i denne gruppe indeholder ikke nye fag, men kommentarer om relationer mellem testede fag mv.

Der er i dette review arbejdet med en tilgang, der ikke umiddelbart skulle give anledning til, at der er mange studier med en sociologisk vinkel. Det er da heller ikke tilfældet. Dette bekræftes af Tabel 3.14:

Er sociologisk vinkler inddraget i forbindelse med testningen?	Antal studier
Nej	53
Ja	7
Uoplyst	1

Tabel 3.14: Inddrages sociologisk vinkling?
(N=61 studier)

Det fremgår at kun 11 % inddrager sociologien i analysen.

Spørgsmålet om kriterier for effektiv testbrug behandles i det følgende. Tabel 3.15 viser omfanget af studier med kriterier hos eleverne, lærerne og i undervisningen:

Hvad ses den effektive testbrug i forhold til?	Antal studier
Lærerkarakteristika	12
Elevkarakteristika	17
Undervisningskarakteristika	32

Tabel 3.15: Hvilke kriterier henviser testbrugen til? (N=61 studier)

20 % af studierne har lagt kriteriet for om testbrugen er effektiv på læreren.

Det der oftest ses på her er forskellige momenter af lærerens undervisningsmæssige praksis.

Det fremgår, at 28 % af studierne anvender kriterier hos eleverne for om testbrugen er effektiv. Indholdsmæssigt er der oftest tale om elevpræstation. Der er enkelte studier hvor andre forhold er kriterier, f.eks. elevernes holdning til læreprocessen.

52 % af studierne bruger kriterier fra undervisningsmiljøet i vurderingen af effektiv testbrug. Indholdsmæssigt arbejder disse studier med meget forskelligartede vinkler som: undervisningsmetodik, læseplan, læringsmiljø, undervisningsprogression.

3.5 De to reviewspørgsmål

I dette systematiske review er der arbejdet med to spørgsmål:

> Hvilken evidens er der for, at grundskolelæreres individ- og klassecentrerede brug af testdatas forbedrer læreres fagdidaktiske tiltag i klasser med almindelige elever? Og
>
> Hvilken evidens er der for, at introduktion af testning influerer på læreres didaktiske beslutninger og elevers læringsadfærd?

Studierne fordeler sig på besvarelse af de to spørgsmål på denne vis:

Svar på reviewspørgsmål	Antal studier
Svar på spørgsmål 1	28
Svar på spørgsmål 2	44

Tabel 3.16: Fordelingen af studierne på de to reviewspørgsmål (N=61 studier; flere kodninger per undersøgelse er mulig)

Der er 72 svar, så 11 studier vedrører begge spørgsmål. 72% af studierne kan bidrage med svar på det andet spørgsmål, medens 46% af studierne kan bidrage til svar på det første spørgsmål.

3.6 Vurdering af forskningskvalitet

Et væsentligt moment i vurderingen af forskningskvalitet er bestemmelsen af den evidens-vægt, som et enkelt studie kan tillægges. Dette er den samlede konklusion med hensyn til evidens vedrørende studiet i dette review, se også Kapitel 7, Appendiks 2 for et eksempel på, hvorledes disse vurderinger er fremkommet i praksis.

Den samlede evidensvægt, evidens D, for de undersøgelser, der indgår i dette systematiske review, er baseret på en samlet vurdering af:

Evidensvægt A: Er vurderingen af om studiet i sig selv er troværdigt som studie. Dette er baseret på en lang række forskningsmetodiske vurderinger af studiets egen kvalitet.

Evidensvægt B: Er vurderingen af om studiet har et forskningsdesign og en analyse der er passende for det aktuelle reviewspørgsmål.

Evidensvægt C: Er vurderingen af om studiet i sit indholdsmæssige fokus (begreber, kontekst, sampling) er relevant for det aktuelle reviewspørgsmål.

Alle evidenser A til D er i forskningskvalitetsvurderingen forbundet med tre niveauer: Lav evidensvægt, medium evidensvægt og høj evidensvægt.

Da alle studier, der er tildelt evidensvægten: lav på evidens C er ekskluderet som ikke relevante under screeningen, bliver der følgende mulige udfaldsrum tilbage i evidens A, B og C.:

Sammenhæng mellem evidensvægtene A,B og C og samlet evidensvægt		
A: H, B: H, C: H	A: M, B: H, C: H	A: L, B: H, C: H
A: H, B: H, C: M	A: M, B: H, C: M	A: L, B: H, C: M
A: H, B: M, C: H	A: M, B: M, C: H	A: L, B: M, C: H
A: H, B: M, C: M	A: M, B: M, C: M	A: L, B: M, C: M
A: H, B: L, C: H	A: M, B: L, C: H	A: L, B: L, C: H
A: H, B: L, C: M	A: M, B: L, C: M	A: L, B: L, C: M

Tabel 3.17: Sammenhænge mellem evidensvægte

I Tabel 1.1 står A, B, C for evidensvægtene A, B, og C. H, M og L står for henholdsvis høj, medium og lav evidens, Dvs. at en celle med følgende kombinatorik: A: H, B: M, C: H står for, at [Evidenvægt] A: har værdien H[øj], at [Evidenvægt] B: har værdien M[edium], og at [Evidenvægt] C: har værdien H[øj]. Og da denne kombinatorik står på en hvid baggrund, har kombinationen samlet fået evidensvægt D: høj, idet den samlede evidensvægt D i tabellen er markeret på følgende måde:

Hvid baggrund: Samlet evidensvægt høj

Lysgrå baggrund: Samlet evidensvægt medium eller høj

Grå baggrund: Samlede evidensvægt medium

Sortgrå baggrund: Samlet evidensvægt lav eller medium

Sort baggrund: Samlet evidensvægt lav

Evidensvægtningen af de inkluderede 61 studier vises i Tabel 3.18. Det fremgår heraf, at 34 % af studierne får samlet evidensvægt høj, og 36 % får samlet evidensvægt medium, medens 30 % får evidensvægten lav. Studierne med den samlede evidensvægt lav inddrages ikke i syntesedannelsen.

	Antal studier		
	Høj Evidens	Middel evidens	Low evidens
Evidens A	26	19	16
Evidens B	23	23	15
Evidens C	29	21	11
Samlet evidens D	**21**	**22**	**18**

Tabel 3.18: Evidensvurdering af inkluderede studier (N= 61 studier)

4 Narrative synteser

4.1 Indledende bemærkninger

I kapitel 3 identificeredes de primærstudier, som indgår i den systematiske syntesedannelse. Skønt der blandt de genbeskrevne studier er eksempler på RCT-undersøgelser (jf., Tabel 3.3), findes de alene som enkeltstående undersøgelser i forhold til de fem hovedrelationer, som syntesedannelsen sker indenfor, jf. nedenfor i afsnit 4.2. Dette udelukker muligheden for at gennemføre systematiske syntetiseringer i form af meta-analyser.

I stedet skal der i det følgende anvendes en procedure, som almindeligvis betegnes som *narrativ syntese i systematiske forskningsreviews* (Popay et al., 2006). Den narrative synteseproces består ifølge denne forståelse af fire elementer, der analytisk præsenteres i en given rækkefølge, men meget vel i den praktiske synteseproces kan indeholde iterative bevægelser mellem de forskellige elementer.

De fire elementer kan kort beskrives på følgende måde:

Første element består i at udvikle en teoretisk model af, hvordan de(n) effekt(er), som er undersøgelsens genstand, virker, hvorfor de(n) virker og for hvem. Man taler undertiden om at opstille en "theory of change" (Weiss, 1998, p. 55), der i Wholey's (1987, p. 78) beskrivelse angiver "the chain of causal assumption that link programme resources, activities, intermediate outcomes and ultimate goals". Teorien anvendes til at tolke reviewets fund og kan være nyttig ved vurderingen af, hvor bredt anvendelige disse fund er.

Andet element består i at udvikle en præliminær syntese. I denne fase er det nødvendigt at organisere de inkluderede studier på en sådan måde, at det er muligt at fastlægge effektens retning og – hvis det er muligt – dens styrke. Samtidig søges der efter mønstre, der også vedrører faktorer der på forskellig måde kan vises at influere på effekten. I denne fase er opgaven at etablere mulige synteser, mens deres robusthed først undersøges i en senere fase.

Tredje element er helliget en gennemgang af de faktorer, der på tværs af studierne kan forklare forskelle i retning og styrke af den undersøgte effekt. Herunder behandler man også spørgsmålet om, hvorfor et fænomen har eller ikke har en effekt, og om der findes særlige forhold, der spiller ind, og som kan forklare, hvordan effekten i en given kontekst styrkes eller svækkes.

Fjerde element er vurderingen af syntesens robusthed. Dette er et komplekst begreb, der noget forenklet kan siges at bestå af tre aspekter.

For det første afhænger en synteses robusthed af de primære studiers *metodologiske kvalitet*. Troværdigheden af en syntese vil både afhænge af kvaliteten heraf og af kvantiteten af den evidensbase, som den bygger på. Hvis primærstudier af ringe metodologisk kvalitet inkluderes i det systematiske review på en ukritisk måde, vil dette påvirke syntesens troværdighed.

Troværdigheden vil for det andet også blive påvirket af de *metoder*, der anvendes i syntesen. Det afhænger bl.a. af hvilke forholdsregler der er brugt til at minimere bias, ved fx at vægte primærstudier af ensartet teknisk kvalitet på en ligelig måde.

Endelig drejer et aspekt sig om, hvorvidt screenere og reviewere har *tilstrækkelig information* til sikkert at inkludere et primærstudie i syntesen. Det kan være et alvorligt problem især i forbindelse med undersøgelsen af effekter knyttet til komplekse forhold, idet det ikke altid af primærstudiet klart fremgår, hvilke sammenhænge de forskellige effekter knytter sig til.

Ved afslutningen af synteseprocessen skal disse aspekter føres sammen og resultere i en overordnet vurdering af styrken af den evidens, hvormed der kan drages konklusioner på basis af en narrativ syntese.

4.2 Narrative synteser på baggrund af den konceptuelle model

Dette afsnit gennemgår de præliminære narrative synteser, som det systematiske review indbyder til. Gennemgangen vil blive struktureret i overensstemmelse med den konceptuelle model, som er blevet præsenteret i afsnit 1.3, og som blev illustreret med hjælp af Figur 1.1, side 32.

Som det vil fremgå, giver det rejste reviewspørgsmål, jf. side 33, anledning ikke blot til én narrativ syntese, men – med udgangspunkt i den konceptuelle model - til en flerhed af narrative synteser. Ved at sammenholde de 43 studier, jf. Tabel 3.18, der samlet set er tildelt forskningskvaliteten "høj" eller "middel", med de følgende lister over de enkeltstudier, der indgår i synteserne, fremgår det, at ét studie *ikke* indgår i de følgende syntesedannelser. Det drejer sig om Mason et al. (2005). Denne undersøgelse er et systematisk review, der har til formål at undersøge effekterne af summative test på gymnasiale læseplaner for billedkunst, lærere og elever. Et af resultaterne ved dette review er, at kvaliteten af forskningen i billedkunst, samt forskningsafrapporteringen er af ringe kvalitet, hvad der gør det vanskeligt at drage forskningsmæssige konklusioner på det foreliggende materiale. Mason et al. (2005) konkluderer, at der er brug for flere primær undersøgelser, der måler effekten af test på faget billedkunst.

Som det allerede er formuleret i afsnit 1.3, kan testning og testdata indvirke på elevens præstation ad tre veje:

Første vej: Når der er foretaget en testning, etableres testdata, dvs. de værdier som retning af items resulterer i. Dette kan give læreren information om både den enkelte elevs præstation og hele klassens præstation. Dette svarer til modellens relation 1. På baggrund af denne information kan læreren beslutte sig for at foretage bestemte didaktiske tiltag. Om læreren gør det og hvilke tiltag, der i givet fald kommer på tale, svarer til modellens relation 2. Har læreren endelig foretaget bestemte didaktiske tiltag på baggrund af information fra testdata, kan dette resultere i, at eleverne lærer (bedre). Dette svarer til relation 3. Samlet set udgør denne "vej", dvs. relation 1, 2 og 3, den mest nærliggende tolkning af udtrykket "pædagogisk brug af test" og kunne præciseres til "pædagogisk brug af testdata".

Anden vej: Når det er bebudet, at klassen skal testes, dvs. at hændelsen 'at der skal gennemføres en test' er annonceret, kan dette influere lærerens didaktiske tiltag. Dette svarer til modellens relation 4. Ligesom ovenfor kan dette resultere i, at eleverne lærer

(bedre). Denne "vej", dvs. relation 4 , udgør ligeledes en mulig tolkning af udtrykket "pædagogisk brug af test". Denne tolkning kan ses i følgende sammenhænge: Uafhængig af om læreren gør sig det bevidst eller ej, bevirker annonceringen af hændelsen 'at der skal gennemføres en test', at læreren (bevidst eller ubevidst) gennemfører bestemte didaktiske tiltag, der via relation 3, indvirker på elevernes læring. Når testen har en officiel eller formel karakter, er det instanser udenfor (over) klasseniveau der beslutter sig for at annoncere og iværksætte hændelsen 'at der skal gennemføres en test' i den forventning, at dette vil influere på lærerens didaktiske praksis, der via relation 3, indvirker på elevernes læring. Man taler om en forventet wash-back effekt. I denne tolkning af "pædagogisk brug af test" vil den følgelig kunne præciseres til "pædagogisk brug af test – virkninger på undervisningen".

Tredje vej: Når det er bebudet, at klassen skal testes, dvs. at hændelsen 'at der skal gennemføres en test' er annonceret, kan dette influere på elevernes læringspraksis. Dette kan ses i to sammenhænge: Uafhængig af om eleven gør sig det bevidst eller ej, bevirker annonceringen af hændelsen 'at der skal gennemføres en test', at eleven ændrer sin læringspraksis. Dette kan både medføre, at elevens indlæring påvirkes. Eller: Instanser udenfor (over) klasseniveau kan beslutte sig for at annoncere og iværksætte hændelsen 'at der skal gennemføres en test' med den forventning, at dette vil influere på elevens læringspraksis. Man taler om en forventet wash-back effekt, der kan vise sig at være både positiv og negativ. Her kan man altså tale om "pædagogisk brug af test" som "pædagogisk brug af test – virkninger på elever".

Endelig bør det ikke glemmes, at modellen også inkluderer en fjerde vej, repræsenteret ved relation 6, som ikke er undersøgt i dette review, jf. side 32. Den vedrører den veletablerede Rosenthal effekt om, at forventes en elev at præstere bedre, så præsterer eleven bedre. I denne sammenhæng udgår forventningen fra læreren. Omvendt: forventes en elev at præstere ringere, så præsterer eleven ringere. Også her udgår forventningen i denne sammenhæng fra læreren.

De konklusioner, som indgår i de enkelte præliminære narrative synteser, som fremstilles i de følgende afsnit 4.2.1, 4.2.2, og 4.2.3, er fremkommet ved følgende procedure: Hvert af de 43 primærstudier, der er tildelt evidensvægten høj og middel, er blevet sammenfattet i et abstract, jf. Kapitel 0 Appendiks 3. Ved hvert abstract er det angivet, hvilken af de 5 relationer, der indgår i den konceptuelle model, jf. Figur 1.1, de bidrager til. Dette bidrag er udformet som en narrativ om, hvad studiet belyser, dvs. hvilke faktorer der har en effekt. Desuden er andre (perifere) forhold af relevans til belysning af denne sammenhæng noteret. Ved de enkelte syntesedannelser er der først taget hensyn til de studier, der har fået evidensvægt høj. Dernæst er studier med evidensvægt medium inddraget. Endelig er der foretaget en sammenfattende tolkning i form af sideordnede og underordnede konklusioner vedrørende den pågældende relation.

Til karakteristisk af en given evidens' styrke er anvendt tre kategorier: *ingen evidens*, dvs. der findes ingen undersøgelser i materialet, der kan bidrage til at belyse spørgmålet, eller undersøgelernes bidrag er ikke konsistente; *svag evidens*, dvs. der findes undersøgelser i materialet, der belyser spørgsmålet, med middel evidensvægt og/eller som undersøgelser med høj og middel evidensvægt, der belyser periferer sider ved spørgsmålet på en konsistent måde; og endelig *stærk evidens*, dvs. der findes undersøgelser i materialet, der belyser spørgsmålet med høj evidensvægt på en konsistent måde.

4.2.1 Pædagogisk brug af testdata: Relation 1, 2 og 3

Først behandles muligheden af syntesedannelse for hver af de tre relationer. Tabel 4.1 giver en oversigt over, hvordan undersøgelser af relevans for relation 1, 2 og 3 fordeler sig på fag.

	Relation 1	Relation 2	Relation 3
Modersmål	4	7	1
Andre sprog	2	6	1
Matematik	3	7	
Naturfag	-	1	-
IT	-	1	1
Samfundsfag	-	1	-
Kunst	-	1	
Ikke specificerede fagområder	1	2	-

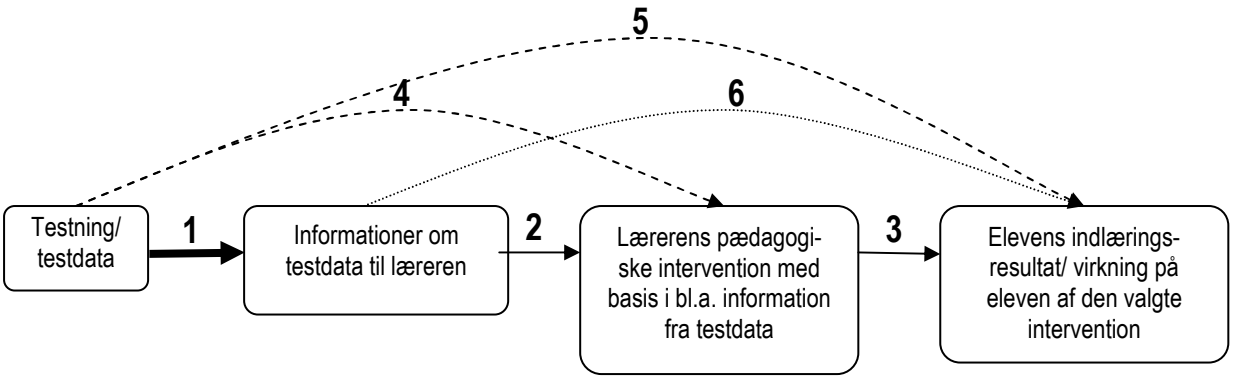
Tabel 4.1: Pædagogisk brug af testdata - fordeling af undersøgelser om fag på relation 1, 2 og 3 (N= 20 studier; flere kodninger per undersøgelse er mulig)

Relation 1

Den første relation, som skal betragtes, fremgår af Figur 4.1.

8 undersøgelser beskæftiger sig med denne relation (Gutkin, 1985; Higgins & Rice, 1991; Jia et al., 2006; Parke et al., 2006; Parker & Picard, 1997; Ryan, 2001; Scott, 2007; Tresch, 2007), heraf 5 med høj evidensvægt.

Relationen drejer sig om, hvilken form for information testdata yder til læreren. Når læreren modtager denne information, tolkes den, og bliver i de gunstige tilfælde til viden om eleven og klassen. Denne viden kan læreren bl.a. lade indgå som én af de baggrundsindsigter, som kan vise sig nyttig i forhold til lærerens pædagogisk/didaktiske overvejelser og beslutninger af den type, som modellen søger at indfange ved hjælp af relation 2.



Figur 4.1: Pædagogisk brug af testdata - relation 1

Ser vi først på spørgsmålet, hvorledes læreren bearbejder testdata til viden om elev og klasse, konstateres det,

- *at ingen af de undersøgelser, der indgår i dette review, beskæftiger sig med lærerens tolkning af de informationer, som testdata yder.*

Vi ser således ikke ud til at kunne sige med evidens, om lærere forstår testdata i overensstemmelse med testkonstruktørernes intentioner eller ej.

Derimod beskæftiger undersøgelserne sig med to forhold af interesse i tilknytning hertil. Det første forhold vedrører spørgsmålet om, hvilke testtyper lærere foretrækker (Higgins & Rice, 1991; Jia et al., 2006), nemlig

- *at lærere foretrækker test, der fremstilles til brug i de individuelle klasser, frem for standardiserede test.*

Der kan ses to begrundelser herfor. Skal testdata anvendes i den daglige undervisning, skal de tage udgangspunkt i forhold, der direkte udspringer af den pågældende klasses curriculum. Er testdata samtidig fremkommet ved, at læreren selv har konstrueret testen, kan testdata lettere integreres i klassens arbejde, fordi læreren bedre forstår, hvilken information der ligger i testdata.

I tilknytning hertil fremhæver en enkelt undersøgelse (Parke et al., 2006),

- *at lærere finder performans-baserede og kognitivt opbyggede test, der stiller krav til eleverne om at udvikle komplekse svar, mest informative.*

Det andet forhold vedrører spørgsmålet om, hvor meget information der ligger i testdata eller de analyser, som foretages heraf (Fuchs & al., 1989; Gutkin, 1985). Generelt gælder, at lærerne mener,

- *at jo mere information de får, des bedre kan de vælge passende undervisningsstrategier og muligvis også designe curriculum og programmer.*

Det fremhæves desuden (Tresch, 2007),

- *at effektiv feedback på test forudsætter en præcis beskrivelse af testen, så læreren ved, hvad der skal måles, og at resultaterne foreligger både på klasseniveau og det individuelle niveau.*

Dette kunne sammenfattes til,

- *at lærere ønsker ejerskab til de test, der anvendes.*

Nogle få undersøgelser har beskæftiget sig med den information læreren får, når der ikke blot foreligger "råtestdata", fx testdata i form af angivelse af et tal eller en tildelt karakter, men også en analyse af de svar, som eleverne har afgivet. En sådan analyse kan være gennemført af forskere eller fagdidaktikere og resultere i en taksonomi over *responstyper*, dvs. en karakteristik af kategorialt forskellige svarmønstre på de stillede items, eller – hvis analysen alene vedrører items, som eleven ikke har løst korrekt - i *fejltyper*. Denne sidste type analyse anvendes især i tilknytning til test i matematik og naturfag, jf. (Duit, 2009). En undersøgelse, hvor fejltyper i matematik inddrages (Ryan, 2001), konkluderer,

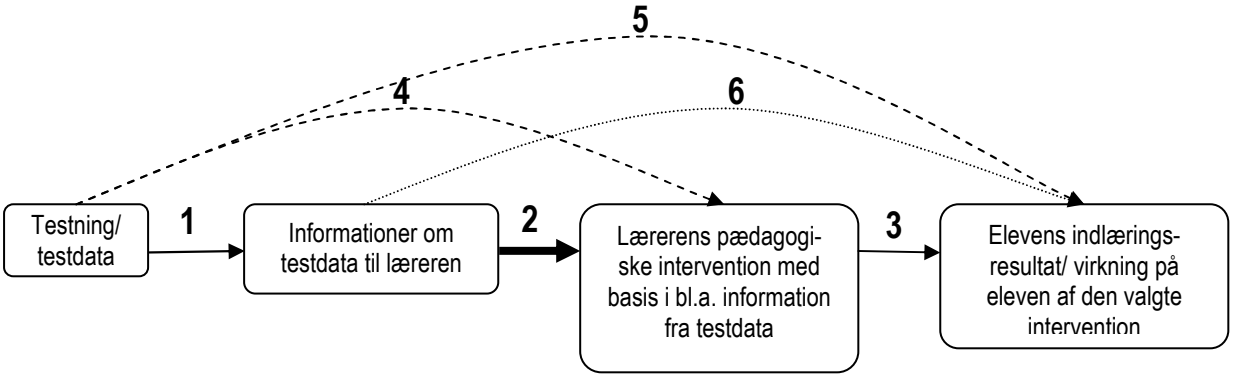
- *at en forskningsmæssig analyse af fejltyper kan etablere baggrund for en lærerviden om mulige fejltyper, der kan være et nyttigt redskab, når læreren vil skabe en mental model af den enkelte elevs forståelse af matematik.*

Men undersøgelsen giver ikke anvisninger på, hvorledes læreren i praksis skal gøre dette. En tolkning kunne være, at undersøgelsen peger på, at testdata ikke bør leveres som "rådata", men bør underkastes specialisttolkning, før de sendes til lærerne.

Endelig viser en enkelt undersøgelse (Parker & Picard, 1997), at data fra en formel test ikke giver retvisende oplysninger om en tosproget elevs færdigheder i matematik.

Relation 2

Den anden relation, som skal betragtes, fremgår af Figur 4.2.



Figur 4.2: Pædagogisk brug af testdata - relation 2

17 undersøgelser beskæftiger sig med denne relation (Alderson & Hamp-Lyons, 1996; Bauer et al., 1990; Birenbaum & Tatsuoka, 1987; Cheng & Rogers, 2004; Fuchs & al., 1989; Harlen & Deakin, 2003; Higgins & Rice, 1991; Howe, 1996; Jia et al., 2006; Kozulin

& Garb, 2004; Mattson, 1989; Parke et al., 2006; Ryan 1994; Ryan, 2001; Scott, 2007; Silis, 2005; Tresch, 2007), heraf 5 med høj evidensvægt.

Relationen drejer sig om, hvilken form for pædagogisk/didaktisk intervention læreren beslutter sig for på baggrund af den information, som testdata yder til læreren.

Her er fire overordnede muligheder; (1) Testdata indgår ikke i lærerens overvejelser ved beslutninger om pædagogisk/didaktisk intervention. (2) Testdata indgår i lærerens overvejelser ved beslutninger om pædagogisk/didaktisk intervention vedrørende klassen, men ikke vedrørende den enkelte elev. (3) Testdata indgår i lærerens overvejelser ved beslutninger om pædagogisk/didaktisk intervention vedrørende den enkelte elev, men ikke vedrørende klassen. (4) Testdata indgår i lærerens overvejelser ved beslutninger om pædagogisk/didaktisk intervention både vedrørende klassen og den enkelte elev.

I de tilfælde, hvor læreren beslutter sig for en pædagogisk/didaktisk intervention, dvs. i tilfældene (2), (3) og (4), kan vi desuden spørge om, hvilken form for pædagogisk/didaktisk intervention læreren gennemfører.

Ser vi først på spørgsmålet, om læreren bearbejder testdata til viden om elev og klasse, konstaterer Alderson & Hamp-Lyons (1996), Bauer et al. (1990), Higgins & Rice (1991) og Jia et al. (2006) om specielt formelle test,

- *at testdata fra formelle test enten slet ikke anvendes eller opfattes som dårligt tilpassede lærerens egne principper for bedømmelse.*

Om dette er en almindelig tendens, kan dette systematiske review ikke udtale sig om. En god grund hertil er de anvendte inklusionskriterier: Vi har søgt efter undersøgelser, hvor læreres pædagogiske brug af testdata behandles, ikke efter undersøgelser, hvor testdata foreligger, men lærere ikke anvender dem pædagogisk.

En række undersøgelser beskæftiger sig med læreres anvendelse af testdata på klasseniveau (Fuchs & al., 1989; Parke et al., 2006; Silis G. 2005; Ryan, 1994; Tresch, 2007). Det er en generel mening,

- *at test betyder meget for undervisningens tilrettelæggelse, især når testen er pædagogisk opbygget, støttet af skolerne, teknisk velfungerende og ikke fokuserer på den enkelte elevs resultat, men på skolens.*

Derimod siger disse undersøgelser ikke noget om, hvordan denne betydning viser sig. Der synes at være en tendens til, at lærere i almindelighed mener, at testdata må være nyttige til tilrettelæggelse af undervisning, men undersøgelserne afdækker ikke, på hvilken måde de er nyttige.

En undersøgelse tilføjer (Tresch, 2007), at lærerne generelt anvender testdata i deres undervisning, men at det kræver, at de har god forståelse af testdata. Test virker godt til at sammenligne klasser og især visse test er gode til at skabe refleksion hos lærerne om deres egne pædagogisk/didaktiske interventioner. Forfatteren anbefaler, at der gives en præcis beskrivelse af testen så læreren ved, hvad der skal måles, at der så vidt muligt opstilles flere sammenligningsvariabler, og at testdata indgår i illustrationer og tabeller.

En anden undersøgelse fremhæver (Silis, 2005),

- *at test især er praktiske for lærerne til planlægning, når de endnu ikke kender eleverne godt.*

Endelig nævner en undersøgelse (Ryan, 1994),

- *at lærere, der anvender testdata som baggrund for deres pædagogiske tiltag, gør det på klasseniveau, ikke på individniveau.*

En række undersøgelser beskæftiger sig ligeledes med anvendelse af testdata på elevniveau, men her i sammenhænge uden for klassen (Birenbaum & Tatsuoka, 1987; Fuchs & al., 1989; Harlen & Deakin 2003; Mattson, 1989; Ryan, 2001; Scott, 2007; Silis, 2005; Tresch, 2007).

Læreres anvendelse af testdata på elevniveau inddrager det helt selvfølgelige, hvis testdata ikke allerede findes i karakterform,

- *at testdata hjælper læreren med at give eleverne karakterer.*

Men det tilføjes, at når en lærer kender en elev bedre, spiller testdata en mindre rolle for karaktergivningen (Mattson, 1989; Silis, 2005).

Lærere anvender også testdata (Silis, 2005; Tresch, 2007) i tilknytning til

- *at føre samtale med elever og forældre.*

Lærere anvender desuden testdata til at få mere viden om eleverne (Silis, 2005; Scott, 2007). På den måde kan data anvendes til at identificeret, hvilke dele af stoffet der ikke blev dækket i den forudgående undervisning, og derved til

- *at belyse problemområder og bidrage til implementering af tiltag, der kan støtte eleven.*

To undersøgelser konstaterer (Fuchs & al., 1989; Tresch, 2007), at testdata fra læsetest medfører,

- *at lærere i det mindste træffer én undervisningsbeslutning per elev.*

Harlen & Deakin (2003) har undersøgt it-baserede test, der registrerer besvarelsen elektronisk. Det konstateres,

- *at it-mediet letter lærerens indsigt i, hvorledes eleven udvikler forståelsen af nyt stof og desuden opgaven, at give feedback på testdata.*

To undersøgelser (Kozulin & Garb, 2004; Ryan, 2001) ser på virkningen af, at læreren får testdata i en særlig form, nemlig med oplysninger om fejltyper. Indsigt i de typer fejl, som en elev begår, kan dels hjælpe læreren til at danne sig en mental model af elevens forståelse af matematik, dels anvendes i fremmedssprogsprogrammer til tilrettelæggelse af særlige individuelle indlæringsstrategier.

Birenbaum & Tatsuoka (1987) har i matematik undersøgt forskellige måder, som læreren kan give feedback på, når et item er besvaret forkert, nemlig ved blot at meddele, at det er svaret forkert, meddele hvad det rigtige svar er, eller ved at meddele, hvilken regneregel der skulle være anvendt og hvori fejlen består. Undersøgelsen har som resultat, at ingen af de tre feedback-meddelelser i sig selv er tilstrækkelige – end ikke den mest informative - til at ændre elevens fejlproces i en følgende test. Undersøgelsen indbyder til den tolkning,

- *at blot retning og korrektion af items ikke i sig selv medfører øget elevlæring.*

En sidste undersøgelse kan, i en venlig fortolkning, vise noget om *forholdet mellem hændelsen 'at der testes' og anvendelsen af testdata,* altså om forhold, der vedrører relation 2 og 4 (Bauer et al., 1990). Den siger,

- *at selve det forhold, at der testes, fuldstændig kan overskygge læreres åbenhed for, at testdata kan anvendes pædagogisk.*

Denne sidste tolkning styrkes, hvis lærerne opfatter testen som high stakes.

Relation 3

Den tredje relation, som skal betragtes, fremgår af Figur 4.3.

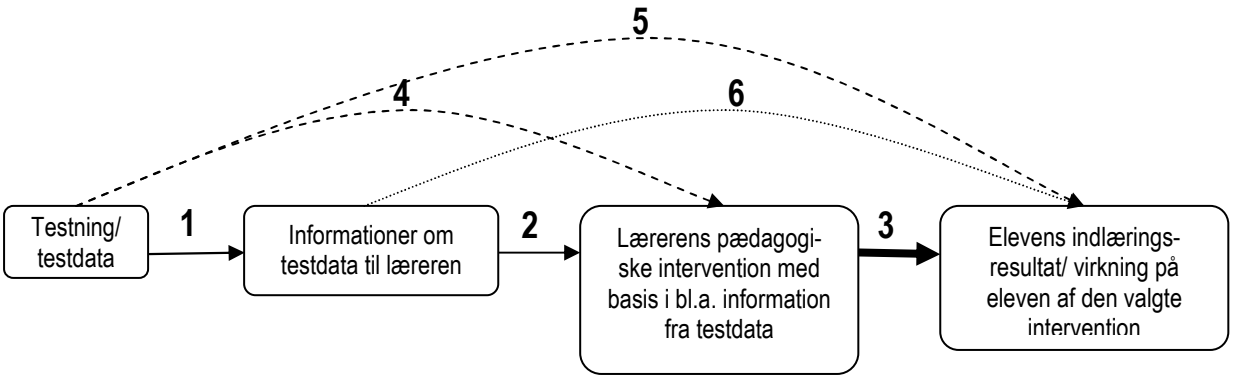
3 undersøgelser beskæftiger sig med denne relation (Gutkin, 1985; Harlen & Deakin, 2003; Kozulin & Garb, 2004), heraf alle med høj evidens.

Relationen drejer sig om effekten på elevlæringen af, at lærere har planlagt deres pædagogiske/didaktiske tiltag på basis af tolkning af information fra testdata.

Alle tre undersøgelser beskæftiger sig med effekten af, at lærere har fået øget information om testdata ud fra yderligere analyser af testdata.

Kozulin & Garb (2004) anvendte et prætest-læring-posttest paradigme ved indlæring af tredje sprog. I lærefasen blev eleverne undervist i strategier, som lettede deres omgang med indlæringsopgaven. De valgte læringsstrategier var individualiserede ud fra prætestdata. Undersøgelsen viste, at fremgangsmåden medfører en signifikant bedre tekstforståelse.

Ifølge en undersøgelse af Gutkin (1985) kunne det ikke påvises, at den øgede information om testdata, som blev tilvejebragt ved hjælp af en superviser, medfører viden, der kan omsættes i bedre elevlæring.



Figur 4.3: Pædagogisk brug af testdata - relation 3

I et systematisk review (Harlen & Deakin, 2003), er det påvist, at feedback på items i opgavesæt om matematik kan tolkes som lærerens pædagogiske intervention på baggrund af testdata. Denne interventionsform lettes ved anvendelse af it-baserede test.

Der er svag evidens for, at dens effekt er bedre, hvor testen er it-baseret, sammenlignet med papir-baserede test.

Pædagogisk brug af testdata: Sammenfatning

Ingen af de studier, der indgår i denne "vej" berører alle tre relationer i samme undersøgelse. Vi har derfor ingen undersøgelser, der *på samme tid* fortæller os, hvilken viden lærere får fra testdata, hvordan de anvender dem til at udforme deres pædagogisk/didaktisk intervention, og hvordan den indlæringsmæssige effekt er af, at læreren har anvendt testdata pædagogisk.

Derimod har vi svag evidens for, at lærere foretrækker så uddybede oplysninger om testdata som muligt, eventuelt ledsaget af respons-/fejltype-analyser. Vi har ligeledes svag evidens for at lærere foretrækker test på klasseniveau, der er tilpasset den aktuelle undervisning, frem for formelle test.

Denne undersøgelse giver en svag evidens for, at lærere anvender test til at planlægge og tilrettelægge undervisningen på klasseniveau. Dette er især tilfældet, hvis læreren ikke kender eleverne godt.

Lærere anvender også test individuelt i forbindelse med karaktergivning, elev- og forældresamtaler, til belysning af faglige problemområder, og pædagogiske tiltag knyttet hertil.

Der er svag evidens for, at det pædagogiske tiltag blot at rette og korrigere fejl i test ikke i sig selv ser ud til at medføre øget elevlæring. Der er også svag evidens for, at pædagogiske tiltag, hvor læreren ud fra testdata underviser i strategier, der letter indlæringsopgaven, giver en øget elevlæring.

Den samlede konklusion er, at den tolkning af "pædagogisk brug af test" som "pædagogisk brug af testdata" er svagt belyst forskningsmæssigt, idet en række ledsagende forhold dog er belyst med svag evidens.

4.2.2 Pædagogisk brug af test – virkninger på undervisningen: Relation 4

Først behandles muligheden af syntesedannelse for relation 4's vedkommende.

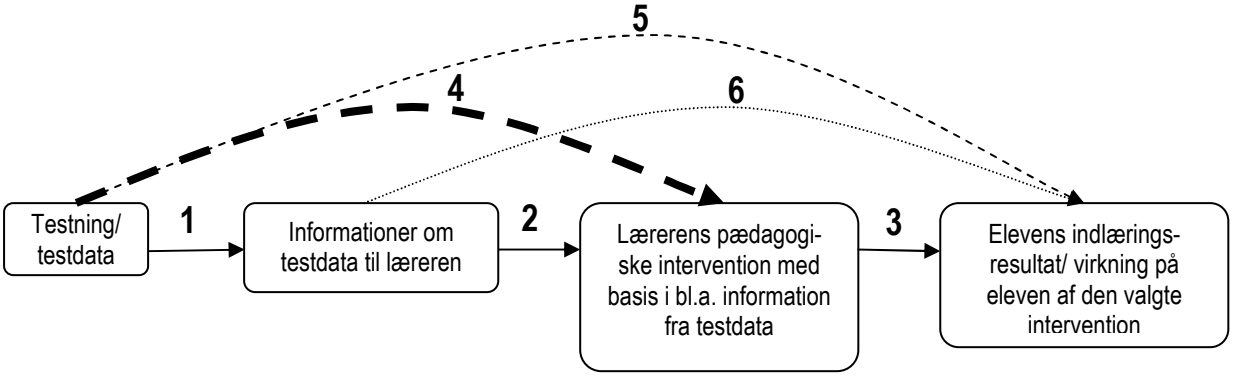
Tabel 4.2 giver en oversigt over, hvordan undersøgelser af relevans for relation 4 fordeler sig på fag.

	Relation 4
Modersmål	6
Andre sprog	8
Matematik	7
Naturfag	6
Samfundsfag	-
Geografi	1
Historie	1
Ikke specificerede fagområder	2

Tabel 4.2: Pædagogisk brug af test – virkninger på undervisningen: Fordeling af undersøgelser om fag på relation 4
(N= 24 studier; flere kodninger per undersøgelse er mulig)

Relation 4

Den fjerde relation, som skal betragtes, fremgår af Figur 4.4.



Figur 4.4: Pædagogisk brug af test – virkninger på undervisningen: Relation 4

24 undersøgelser beskæftiger sig med denne relation (Alderson & Hamp-Lyons, 1996; Andrews et al., 2002; Bauer, 1990; Boesen, 2006; Chen et al., 2002; Cheng & Curtis, 2004; Danmarks Evalueringsinstitut, 2002; Ferman, 2004; Firestone et al., 2004; Jia et al., 2006; Lane et al., 2002; Goldberg & Roswell, 1999; Harlen & Deakin, 2002; Luxia, 2007; NFS, 1992; Parke et al., 2006; Shohamy et al., 1996; Smart, 2007; Smith et al., 1992; Stecher et al., 2004; Stone & Lane, 2003; Sturman, 2003; Wall & Alderson, 1992; Watanabe, 2004), heraf 14 med høj evidensvægt.

Relationen drejer sig om den virkning, som hændelsen 'at der skal gennemføres en test' har på lærerens pædagogiske/didaktiske tiltag. Om lærere er sig denne virkning bevidst eller ej, er i første omgang ikke så interessant. Mere afgørende er, at beslutningen om, at der skal testes, i denne sammenhæng kommer fra et ledelsesmæssigt højere niveau i

uddannelsessystemet end læreren i klassen, nemlig enten fra skolelederen, regionale organer, og/eller centrale organer, herunder undervisningsministerier/-ministre. En sådan beslutning skyldes ofte forestillingen om en "test-drevet reform" (*assessment-driven reform*).

En "test-drevet reform" bygger på den antagelse, at curriculum og undervisning kan indføres og ændres gennem test-programmer. Adskillige undersøgelser i dette systematisk review evaluerer bestræbelsen på at anvende test til netop at reformere uddannelsessystemet eller gennemføre en undervisningsreform. Forskningen om denne effekt betegnes almindeligvis som en undersøgelse af wash-back (eller back-wash) effekten af introduktion af en test. Det spørgsmål som undersøges er altså, om introduktionen af en test vil medføre, at lærere forud for testningen ændrer curriculum og undervisning i overensstemmelse med de politiske ønsker i den reform, som testen tester i? Når testen er gennemført får man samtidig som en sidegevinst et indtryk af, i hvilken udstrækning intentionen med reformarbejdet genspejler sig i elevlæringen.

To muligheder kan tænkes, nemlig (1) at introduktion af en test ingen forskel gør og (2) at den gør en forskel. I det sidste tilfælde kan der yderligere skelnes mellem, om wash-back effekten er (a) positiv, dvs. i overensstemmelse med intenderede effekter, (b) negativ, dvs. der opstår uønskede effekter, som modarbejder intenderede effekter, og (c) uintenderet, dvs. medfører effekter, som ikke er planlagt. De kan både være positive og negative.

6 undersøgelser rapporterer om positiv wash-back effekt på lærerens undervisning i klasselokalet (Bauer, 1990; Danmarks Evalueringsinstitut, 2002; Goldberg & Roswell, 1999; Lane et al., 2002; Parke et al., 2006; Stone & Lane, 2003). De samme undersøgelser rummer også observationer af både uønskede og uintenderede virkninger af den samme test. Dette viser overordnet,

- *at relationen mellem testning og undervisning er kompleks og i høj grad afhængig af kontekstuelle faktorer.*

En afgørende kontekstuel faktor er de reelle eller forestillede følger, som testdata har eller kan have for de involverede parter, i denne sammenhæng lærerne og eleverne. Man skelner mellem to polære muligheder, nemlig *high stakes* og *low stakes* test, men principielt er der tale om mulige, graduelle positioner mellem de to poler. En *high stakes* test har eller opfattes at have vigtige konsekvenser for lærerens og/eller elevens fremtid, fx ved at være afgørende for forfremmelse eller ved at regulere adgangen til en attraktiv uddannelse, mens en *low stakes* test ikke har vigtige konsekvenser for lærerens eller elevens fremtid. I stedet kan de fx give læreren indsigt i en klasses aktuelle faglige niveau eller give en elev den oplysning, at visse dele af et fag ikke er forstået.

Alle 6 undersøgelser rapporterer om,

- *at der er positive wash-back effekter - fra svagere til stærkere - med hensyn til testenes indflydelse på lærerens undervisning i intenderet retning.*

To undersøgelser er overvejende positive. Lane et.al. (2002) undersøgte effekten af Maryland School Performance Assessment Program (MSPAP) og the Maryland Learning Outcomes (MLOs) vedrørende undervisningen og bedømmelsespraksis i matematik, professionel udvikling og elevlæring. De konkluderer, at når lærere rapporterer om ændringer i undervisningspraksis, er det mere sandsynligt, at de ikke blot forbereder til MSPAP-

testen, men har foretaget virkelige ændringer, der forbedrer elevernes forståelse af matematik. I en undersøgelse (Stone & Lane, 2003) året efter af sammenhængen mellem MSPAP-programmet og den gennemførte undervisningspraksis i perioden fra 1993 til 1998, fandt man en parallel udvikling mellem undervisning i overensstemmelse med reformen og forbedring af testresultater.

De øvrige 4 undersøgelser hævder på den ene side, at effekten er positiv, men samtidig,

- *at effekten ikke er entydig positiv, idet lærerne var skeptiske til negative over for de anvendte test, hvis de ikke selv havde været involveret i at udvikle dem.*

Bauer et. al. (1990) rapporterer om introduktionen af et nyt undervisningsprogram og af the New York State program evaluation tests, PETs, til at evaluere det. Lærerne var overvejende negativt indstillede og mente ikke selv, at PETs ændrede deres curriculum. De opfattede ikke PETs som en evaluering af et program, men som en test på individniveau, der både var high stakes for lærere og elever.

Danmarks Evalueringsinstitut (2002) fandt, at lærerne opfattede test som en kontrol af undervisningen. Lærerne praktiserede overvejende helklasseundervisning i de fag, som blev afsluttet med test, hvor eleverne også var mindre aktive end i fag uden test.

Goldberg & Roswell (1999) undersøgte læreres instruktion og anvendelse af den ovenfor nævnte, specifikke test, the Maryland School Performance Assessment Program (MSPAP), der indebar test i form af praktiske opgaver. Det noteres, at lærere kun havde tilegnet sig færdigheden i at undervise i de praktiske opgaver på en overfladisk og ufuldstændig måde.

Parke et.al. (2006) undersøgte også wash-back effekten af MSPAP-testen, der refererer til specifikke standarder formuleret i the Maryland Learning Outcomes (MLO). Forfatterne fandt, at der var diskrepans mellem lærernes opfattelse af egen praksis og den observerede, aktuelle praksis.

I denne sammenhæng bør også Wall & Alderson (1992) nævnes. De fandt, at der var en forskel mellem, hvad lærerne hævdede om deres undervisning, og den måde de faktisk gennemførte den på.

De to undersøgelser peger på,

- *at det kan forekomme, at lærere rapporter om ændret undervisningsform og indhold, som ikke kan konstateres, når undervisningen bliver observeret.*

16 studier rapporterer (Alderson & Hamp-Lyons, 1996; Andrews et al., 2002; Boesen, 2006; Chen et al., 2002; Cheng et al., 2004; Ferman, 2004; Firestone et al., 2004; Harlen & Deakin, 2002; Jia et al., 2006; Luxia, 2007; NFS, 1992; Shohamy et al., 1996; Stecher et al., 2004; Sturman, 2003; Wall & Alderson, 1992; Watanabe, 2004),

- *at en test eller en test-plan ikke har nogen effekt, eller har en overvejende negativ effekt.*

Effekterne fordeler sig på følgende tre områder: (a) indvirkning på curriculum, (b) indvirkning på tid allokeret til det/de fag, der skal testes i, (c) og indvirkning på den måde, der undervises.

En række undersøgelser (Alderson & Hamp-Lyons, 1996; Boesen, 2006; Harlen & Deakin, 2002; Chen et al., 2002; Ferman, 2004; Jia et al., 2006; NFS, 1992; Watanabe, 2004) fremdrager,

- *at introduktion af test medfører et indsnævret eller fordrejet curriculum,*

idet læreren afgrænser undervisningen til at dreje sig om det, der skal testes i (*"teaching to the test")*.

Firestone et al. (2004) hævder, at den amerikanske måde at udvikle standarder og ansvarlighed på, fokuserer for meget på bedømmelse og sanktioner og for lidt på at øge lærernes viden om fag og fagdidaktik. Stone & Lane (2003) påpeger,

- *at sammenligner man lavt præsterende skoler med højt præsterende skoler, anvender lærere mere tid på test i lavt præsterende skoler end i højt præsterende.*

Ligeledes kan virkningen være (Harlen & Deakin, 2002; Chen et al., 2002; Danmarks Evalueringsinstitut, 2002; Ferman, 2004; Smart, 2007; Stecher et al., 2004; Sturman, 2003),

- *at læreren anvender mere tid på de fag, der testes i, på bekostning af de fag, der ikke testes i.*

Endelig finder man (Alderson & Hamp-Lyons, 1996; Andrews et al., 2002; Harlen & Deakin, 2002; Chen et al., 2002; Ferman, 2004; Shohamy et al., 1996),

- *at lærerens undervisningsmetoder begrænser sig til at træne til testen,*
- *at undervisningen bliver mindre interaktiv, og*
- *at undervisningen i sprog fremtræder mere unaturlig,*

hvorved faglige tankegange forsimples, faktaviden og mekaniske færdigheder betones på bekostning af kreative og æstetiske perspektiver.

Disse fremgangsmåder kan være meget effektive til at hjælpe eleverne til at bestå testen, *også* i de tilfælde, hvor eleven ikke forstår eller ikke anvender de mere avancerede ræsonnementer, som det er intentionen, at testen skal teste i (Harlen & Deakin, 2002).

Pædagogisk brug af test – virkninger på undervisningen: Sammenfatning

Det er den "vej", som forskningsmæssigt er bedst belyst.

Undersøgelserne behandler wash-back effekten af at introducere formelle test. Overordnet viser de, at relationen mellem testning og undervisning er kompleks og i høj grad afhængig af kontekstuelle faktorer.

Der er med svag evidens påvist positive wash-back effekter – fra svagere til stærkere - med hensyn til testenes indflydelse på lærerens undervisning i intenderet retning. Der er ligeledes påvist svag evidens for, at der er forskel på læreres opfattelse af ændring af undervisningsform og indhold, og de aktuelle observerede ændringer.

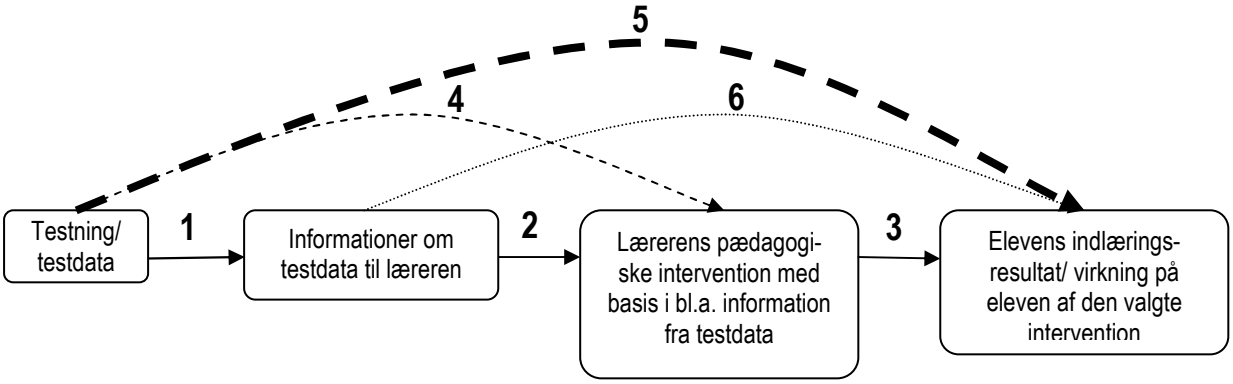
Undersøgelserne viser også betydelig negative wash-back effekter af introduktion af formelle test. Effekterne fordeler sig på følgende tre områder: (a) indsnævret eller fordrejet curriculum, idet faglige tankegange forsimples, faktaviden og mekaniske fær-

digheder betones på bekostning af kreative og æstetiske perspektiver (b) undervisningstid allokeres til det/de fag, der skal testes i, på bekostning af de fag, der ikke testes i, og (c) undervisningen kan forfalde til træning til test og udenadslæren.

En samlet vurdering viser, at fordelene ved introduktion af formelle test kan opvejes af de negative virkninger, som disse test har især for de svagere præsterende elevers vedkommende.

4.2.3 Pædagogisk brug af test – virkninger på eleven: Relation 5

Den femte relation, som skal betragtes, fremgår af Figur 4.5.



Figur 4.5: Pædagogisk brug af test – virkninger på eleven: Relation 5

13 undersøgelser beskæftiger sig med denne relation (Alderson & Hamp-Lyons, 1996; Andrews et al., 2002; Danmarks Evalueringsinstitut, 2002; Doran, 2002; Ferman, 2004; Harlen & Deakin, 2002; Henderson et al., 2007; NFS, 1992; Lane et al., 2002; Parke et al., 2006; Scott, 2007; Shannon, 1980; Stone & Lane, 2003), heraf 7 med høj evidensvægt.

Tabel 4.3 giver en oversigt over, hvordan undersøgelser af relevans for relation 5 fordeler sig på fag.

Relationen drejer sig om den virkning, som hændelsen 'at der skal gennemføres en test' har på eleverne. Om de er sig denne virkning bevidst eller ej, er i første omgang ikke så interessant. Derimod er flere kontekstuelle forhold relevante: De reelle eller forestillede følger, som testdata har eller kan have for den involverede elev, dvs. om testen opfattes som eller er en *high stakes* eller *low stakes* test, elevens socioøkonomiske status, og elevens opfattelse af egen faglig dygtighed.

I modsætning til de påvirkninger, som er medieret af lærerens undervisning gennem relationerne 1, 2 og 3, er de her omtalte virkninger en direkte følge af annonceringen af, at der skal testes. Disse virkninger kan ligge forud for testningen, være aktive under testen, og have følger, som ikke optræder, hvor der ikke testes. Virkningerne fordeler sig over fem forhold: følelsesreaktioner, læringsstil, motivation, selvværd og effekter fra elevens testresultater.

	Relation 5
Modersmål	6
Andre sprog	2
Matematik	6
Naturfag	3
Samfundsfag	1
Geografi	1
Historie	1
Ikke specificerede fagområder	1

Tabel 4.3: Pædagogisk brug af test – virkninger på eleven: Fordeling af undersøgelser om fag på relation 5
(N= 12 studier; flere kodninger per undersøgelse er mulig)

Sammenhængen mellem de fem forhold kan opfattes på følgende måde: Det annonceres, at der testes. Det giver anledning til følelsesreaktioner hos eleverne. Derefter går de i gang med at forberede testen og vælger en læringsstil. Både før og efter testen påvirkes elevernes motivation. Når testen er gennemført, foreligger effekten af, at der testes, dels på testdata og dels på elevernes selvværd. Disse fem forhold skal nu kommenteres mere detaljeret.

Når hændelsen 'at der skal gennemføres en test' annonceres, kan det hos nogle elever give anledning til en følelsesreaktion.

Alderson & Hamp-Lyons (1996) rapporterer,

- *at i klasser, hvor en high stakes test afslutter undervisningen, er der mindre latter og munterhed,*

mens Ferman (2004) konstaterer,

- *at 3 ud af 4 eleverne rapporterer følelsen af angst og ængstelse for testresultatet.*

Det er påvist især for bogligt svage elevers vedkommende. Harlen & Deakin (2002) finder specielt,

- *at piger i højere grad end drenge viser testangst.*

I denne sammenhæng giver Lane et al., (2002) et interessant perspektiv. Hvis en test kun har high stakes konsekvenser for skolen, ikke for den individuelle elev, fandt man, at angstniveauet blev lavere både for lærere og elever uden at svække elevernes ønske om at gøre deres bedste.

Når det er annonceret, at der skal testes, kan det ligeledes få indflydelse på den læringsstil, som eleverne udviser før prøven. Den overordnede tendens er,

- *at eleverne i øget udstrækning benytter sig af udenadslæren og memorering.*

Ferman (2002) fandt, at især bogligt svage elever anvendte cue cards, der hjælper til at memorere forud udarbejdede sætninger. NFS (1992) nævner, at centrale, og formelle test begrænser elevernes tænkning, og Danmarks Evalueringsinstitut (2002) konstaterer, at eleverne bliver i mindre grad udsat for varierende undervisning og metoder, der påkalder høj elevdeltagelse. Endelig noterer Harlen & Deakin (2002), at faktaviden betones på bekostning af elevaktivitet og kreativitet.

En undersøgelse af Andrews et al. (2002), der knytter sig til dette punkt, finder,

- *at ganske vist scorer eleverne højt nogle år efter, at en test er blevet indført, men de øgede elevresultater skyldes memoreren af tekst.*

Dette viser studiet ved at følge mundtlige eksamener.

Flere studier finder,

- *at elevernes motivation øges før testen.*

Danmarks Evalueringsinstitut (2002) taler om, at test synes at motivere og disciplinere både lærere og elever, ligesom Ferman (2004) påviser øget motivation for at lære sproglige færdighed, når de studerende skal forberede testen.

I den samme undersøgelse (Ferman, 2004) konstateres det imidlertid også, at da de studerende tilsyneladende har nemt ved at "snyde" til eksamen, har det en negativ indvirkning på motivationen for at lære. De studerende bekræfter, at de kopierer opgaver fra andre studerende, downloader resumeer fra nettet og ser film baseret på bøger i stedet for at læse dem.

Elevernes motivation påvirkes af det testresultat, de opnår. Shannon (1980) fandt,

- *at testresultatet giver en positiv holdning til læsning blandt gode læsere, men giver en negativ holdning hos svage læsere.*

Når testen er gennemført, foreligger testdata. Doran (2002) og Henderson et al. (2007) kunne ikke påvise nogen effekt af introduktion af test, mens Andrews et al. (2002), Lane & al. (2002), Parke et al., (2006) og Stone & Lane (2003) rapporterer om øget elevlæring. Lane & al. (2002) og Parke et al., (2006) konstaterer samtidig, at elevers testresultater har en signifikant stigning på skoler, der rapporterer øget brug af testvenlige klasserumsaktiviteter (*"teaching to the test"*).

Endelig har testresultater en markant indvirkning på elevernes selvværd (Scott, 2007). Harlen & Deakin (2002) har i et systematisk review konkluderet: Studier vurderet til at have høj evidensvægt viser,

- *at efter introduktionen af* the National Curriculum Tests *i England havde lavt præsterende elever lavere selvværd end højtpræsterende elever. Før introduktionen af disse test var der ingen korrelation mellem selvværd og præstationer (*achievement*).*

Studier med medium evidensvægt bekræfter tilsvarende,

- *at de statsstyrede test i USA bevirker lavere selvværd for elever i risikozonen.*

Pædagogisk brug af test – virkninger på eleven: Sammenfatning

Der er svag evidens for, at elevernes testresultater stiger ved indførelse af test, men først efter nogle år. Når en test annonceres, er der svag evidens for, at det kan udløse følelsesreaktioner som nervøsitet og angst, at eleven forbereder sig ved at lære udenad og memorere sætninger. For bedre præsterende elever stiger motivation, mens svagere præsterende taber modet. Der er stærk evidens for, at det testresultat, som eleven får ved testen, kan virke ind på fremtidig motivation og selvværd.

4.3 Retning og styrke af de undersøgte effekter

Den narrative synteses tredje element består i en gennemgang af de faktorer, der på tværs af studierne kan forklare forskelle i retning og styrke af den undersøgte effekt. Herunder behandler man også spørgsmålet om, hvorfor et fænomen har eller ikke har en effekt, og om der findes særlige forhold, der spiller ind, og som kan forklare, hvordan effekten i en given kontekst styrkes eller svækkes.

I afsnit 1.3 er allerede hypotetisk formuleret, at påvirkningens retning ad tre veje kan føre til indvirkning på elevens læring. Samtidig er det angivet, at denne virkningssammenhæng sker i en kontekst. I det følgende skal de to aspekter, hhv. påvirkningens retning og styrke, og den kontekstuelle sammenhæng, behandles hver for sig.

4.3.1 Påvirkningens retning og styrke

Afsnit 4.2 gennemgik i detaljer de tre veje, som ved brug af test kunne føre til påvirkning af elevens læring. Det blev undersøgt, i hvilken udstrækning læreren anvendte testdata til pædagogisk intervention, og om netop denne brug førte til elevlæring. Det blev ligeledes undersøgt, om annonceringen af hændelsen 'at der skal gennemføres en test' gennem lærerens pædagogiske reaktion herpå, førte til elevlæring. Og endelig blev det undersøgt, om annonceringen af hændelsen 'at der skal gennemføres en test' direkte indvirker på elevens læring, og da på hvilken måde.

Påvirkningens retning er derfor fra testning/testdata til elevlæring.

Kun et sted i den underliggende, konceptuelle model, jf. Figur 1.1, er der usikkerhed om påvirkningens retning. I afsnit 4.2.3 blev det nævnt, at eleven påvirkes både af annonceringen af hændelsen 'at der skal gennemføres en test' og af information om testresultatet. Forud for gennemførelsen af testen påvirkes elevens motivation, og efter gennemførelsen af testen påvirkes elevens motivation også, denne gang som reaktion på testresultatet. Hvis vi tager det forhold alvorligt, at elevens møde med testen også medfører en læreproces, der vedrører elevens fremtidige lyst til at beskæftige sig med det emne, som testen tester i, må det antages, at motivationen før og efter testen interfererer på en ikke afklaret måde. Kan resultatet fra en undersøgelse (Shannon,1980) imidlertid tages for pålydende, tyder det dog på, at den efterfølgende lyst til/holdning til det testede emne, primært påvirkes af oplysningen om testresultatet og ikke af annonceringen om, 'at der skal gennemføres en test'. Den motivation, som eleven har før testen gennemføres, "overlever" så at sige ikke effekten på elevens fremtidige motivation på basis af oplysningen om testresultatet. Vi kan derfor fastholde, at påvirkningens retning er fra testning/testdata til elevlæring.

Styrken af påvirkningen er det imidlertid vanskelige at bedømme. Det er allerede i de tre sammenfatninger omtalt i afsnit 4.2 ovenfor konstateret, at de relationelle forhold i nogle tilfælde fremtræder med ingen eller svag evidensvægt. Det betyder ikke, at det ikke godt kan være tilfældet, at der her foreligger en påvirkning af en vis styrke, men at dette på baggrund af den eksisterende forskning endnu savner klare beviser.

En anden betragtning bør imidlertid bringes ind. I den her gennemførte undersøgelse har det været påstanden, at eleven påvirkes ad tre veje. Det har ligeledes været påstået, at de alle på den ene eller anden måde bidrager til elevens læring.

Men det har endnu ikke været undersøgt, om de alle trækker i samme retning, dvs. om alle påvirkninger af test/testning kan "adderes" eller om nogle skal "subtraheres". Det kan fx meget vel tænkes, at selvom læreren anvender testdata fra en tidligere test til at tilrettelægge og tilpasse undervisningen i klassen, modvirker annonceringen af 'at der skal gennemføres en test' elevens koncentration om opgaven (eleven bliver nervøs) og motivation til at anstrenge sig (forventer ikke at klare testen godt).

Som det blev noteret ovenfor viser én undersøgelse noget om forholdet mellem hændelsen 'at der skal gennemføres en test' og anvendelsen af testdata, altså om forhold, der vedrører relation 2 og 4 (Bauer et al., 1990). Den siger, at selve det forhold, at der testes, fuldstændig kan overskygge læreres åbenhed for, at testdata kan anvendes pædagogisk.

Videre må det ikke glemmes, at også Rosenthal effekten, der ikke her er undersøgt, men forudsættes at spille ind, også kan virke både positivt og negativt ind på elevens læring.

Vi må derfor her minde om den på side 68 fremsatte konstatering, at relationen mellem testning og undervisning er kompleks og i høj grad afhængig af kontekstuelle faktorer.

4.3.2 Den kontekstuelle sammenhæng

I forskningskortlægningens afsnit 3.1, 3.2, 3.3 og 3.4 er noteret mulige kontekster, som kunne tænkes at være af betydning for vurderingen af påvirkningens retning og styrke. Tabel 3.1, Tabel 3.4, Tabel 3.5, Tabel 3.8 og Tabel 3.13 opregner henholdsvis de lande, hvor studierne er gennemført, de overordnede testformål, hvad testen angår, hvem der scorer testen, og de skolefag, som er blevet undersøgt. Flere kontekstuelle forhold kan tænkes, jf. hele kapitlet 3. Kun én undersøgelse har en komparativ karakter, der tillader en vurdering af det kulturelle klimas betydning for holdninger til test. Cheng et al. (2004) viser, at eksempelvis canadiske lærere har en mere negativ holdning til test end kinesiske lærere. Dette henføres til forskellige undervisningskulturer.

Dertil kommer de kontekster, som gennemgangen af de kortlagte undersøgelser fremdrager. Det er allerede nævnt side 71, at de reelle eller forestillede følger, som testdata har eller kan have for den involverede elev, dvs. om testen opfattes som eller er en *high stakes* eller *low stakes* test, kan være en afgørende faktor. Ligeledes gælder det elevens socioøkonomiske status, og elevens opfattelse af egen faglige dygtighed. Kun en enkelt undersøgelse henviser direkte til betydningen af elevens etniske baggrund (Parker & Picard, 1997). I alle tre tilfælde gælder det, at high stakes, lav socioøkonomisk status og lav tiltro til egen dygtighed alle indvirker negativt på elevens testpræstation, motivation og selvværd.

Endelig bør det noteres, at baggrundsfaktorer som lærerens køn, alder, socioøkonomiske baggrund og etnicitet i det store og hele ingen rolle spiller i de kortlagte undersøgelser.

4.4 De narrative syntesers robusthed

I den narrative synteses fjerde element forsøger man at vurdere de etablerede syntesers robusthed. Her indgår, som omtalt ovenfor tre aspekter: de primære studiers *metodologiske kvalitet*; de *metoder*, der anvendes i den narrative syntese; og den grad af *information* om primærstudierne, der har tilladt inklusion i det systematiske review. De tre aspekter skal i det følgende omtales hver for sig. Det skal dog bemærkes, at der allerede i afsnit 3.6 er foretaget en samlet kvalitetsvurdering af de kortlagte undersøgelser med speciel omtale af studiernes rapporteringskvalitet og bidrag til evidens.

4.4.1 De primære studiers metodologiske kvalitet

Det første aspekt drejer sig om de primære studiers metodologiske kvalitet. Tabel 3.3, registrerer det design, som de kortlagte studier har anvendt. I denne forskningskortlægning indgår studier med både "høj", "medium" og "lav" evidensvægt, Tabel 3.18. Det enkelte studies overordnede evidensvægt er som tidligere beskrevet baseret på en samlet vurdering af forskningens og konklusionernes troværdighed, relevans af studiets sigte og hensigtsmæssighed af det valgte undersøgelsesdesign og analyse til at besvare reviewspørgsmålet, jf. Tabel 3.17. Det skal nævnes, at studiernes overordnede evidensvægt derfor godt kan være højere eller lavere end de enkelte aspekter i vurderingen. Renses oversigten over anvendte design for undersøgelser med evidensvægt "lav", giver det følgende fordeling af anvendte forskningsdesign for de studier, der indgår i de narrative synteser, jf. Tabel 4.4Tabel 4.4: Fordeling af forskningsdesign som er anvendt i de narrative syntese.

	High	Medium	Total
Eksperiment med ikke-randomisert fordeling på grupper	3	1	4
En gruppe før-etter-test	2	1	3
En gruppe kun etter-test	1	1	2
Cohort studier	2	4	6
Case-control studier	0	1	1
Tverrsnitsstudier	5	7	12
Menings(=views) studier	9	10	19
Etnografiske studier	1	5	6
Systematisk review	1	2	3
Case studier	3	2	5
Metodologiske studier	1	0	1
Sekundær dataanalyse	1	0	1
Dokumentstudier	2	3	5

Tabel 4.4: Fordeling af forskningsdesign som er anvendt i de narrative synteser (N = 43 undersøgelser; flere kodninger per undersøgelse er mulig)

Af de totalt 61 studier som indgår i undersøgelsen, er 43 (70 %) tilbage, når studier med lav evidensvægt er udeladt. Af disse 43 er 21 vurderet til at have høj evidensvægt, mens 22 er vurderet til at repræsentere medium evidensvægt. De præliminære narrative synteser er, hvor dette har vært mulig, baseret hovedsagelig på studier med høj evidensvægt, men i og med at nogen af de undersøgte relationer i modellen kun er belyst af et fåtal undersøgelser, har det ikke altid været nyttig at skelne mellem disse evidensvægte i syntesen, jf. **.

De vurderinger, der allerede er fremsat i Afsnit 3.1.1, er relevante i tilknytning til dette første aspekt vedrørende studiernes metodologiske kvalitet. Dertil kommer følgende betragtning: Rieper & Foss Hansen (2007, 79, fig. 7.1) har på baggrund af Petticrew & Roberts (2003, 2006, p. 60) opstillet en evidenstypologi om sammenhængen mellem forskningsspørgsmål og forskningsdesign. Af denne typologi fremgår det, at hvor reviewspørgsmålet drejer sig om effektstudier, tildeles randomiserede, kontrollerede eksperimenter den største evidensvægt, efterfulgt af cohort studier og quasi-eksperimentelle undersøgelser (repræsenteret ved de tre** øverste kategorier i Tabel 4.4). Vi kan først slå fast, at af de kortlagte undersøgelser repræsenterer ingen randomiserede, kontrollerede eksperimenter**. Der er 6 cohort studier og 9 med quasi-eksperimentelt design, og korrigeret for dobbeltmarkering i Tabel 4.3 er 13** tilbage af de i alt 43 undersøgelser (30 %), som kan siges at have middel evidensvægt i Rieper & Foss Hansens evidenstypologi.

Det er i denne sammenhæng interessant at registrere nogle uligheder mellem de undersøgelser, som er inkluderet i dette systematiske review, sammenlignet med et andet review som Clearinghouse afsluttede i 2008 om effekter af lærerkompetencer (Nordenbo et al., 2008). Begge reviews havde som udgangspunkt, at de ønskede at studere effekt i en eller anden form. I reviewet om læreres kompetence blev nogle flere studier inkluderet (70 studier), en noget større andel af de inkluderede studier blev vurderet til at have middel eller høj evidensvægt (79 %), og en langt højere andel af studierne (69 %) havde et quasi-eksperimentelt design. Til trods for dette er en langt større andel af de inkluderede undersøgelser i det foreliggende review blevet bedømt til at have høj evidensvægt. I reviewet om lærerkompetencer blev ca. 20 % av de inkluderede studier bedømt til at have høj evidensvægt, mens det i det foreliggende review har ca. halvdelen af studierne fået en tilsvarende vurdering. Dette fortjener en nærmere diskussion.

For det første er det værd at lægge mærke til, hvilke design som i større grad er inkluderet i dette reviewet sammenlignet med reviewet om lærerkompetencer. Den store forskel er det store antal *meningsstudier*, *som er inkluderet*. Dette er studier, hvor forskerne forsøger at forstå et fænomen ud fra socialt betingede normer og værdier hos en gruppe, en kultur eller et samfund. Eksempler på dette kan være studier som forsøger at kortlægge og/eller forstå et fænomen ud fra en bestemt ideologisk, filosofisk eller sociologisk begrundet position. Typiske eksempler på sådanne studier er forskellige former for undersøgelser af læreres opfattelser og meninger om virkninger af test på undervisningen og elevens læring.

Alt i alt siger dette, at den metodologiske kvalitet for inkluderede studier kan bedømmes som relativt lav, hvis vi lægger de kriterier for effektstudier fremsat af Rieper & Foss Hansen (2007) til grund. Imidlertid er der en væsensforskel med hensyn til, hvordan effektstudier er ulige vægtlagt i de to reviews, som vi her sammenligner. I reviewet om lærerkompetencer var studiet af effekter en *nødvendig forudsætning* for at blive klassificeret som havende høj evidensvægt. Også i dette review skal søgning og screening i udgangspunktet sikre, at referencer i databaserne, som dokumenterer eventuelle effekter af test på for eksempel elevers læringsadfærd, er inkluderet, men inklusionskriterierne og reviewspørgsmålet definerer ikke dette som en nødvendig forudsætning for at blive regnet som havende høj evidensvægt. Det er derfor naturligt, at den evidensvægt som reviewgruppens medlemmer har tildelt studierne, ikke hænger sammen med de kriterier, der er fremsat af Rieper & Foss Hansen. Generelt er de eksperimentelle design anvendt i mindre grad inden for uddannelsesforskning[6], hvad der kan forklare den store forekomst af andre design i dette review. Det er derfor nødvendigt at påpege, at sammensætningen af design i de studier, som er inkluderet, har som konsekvens, at vi i dette systematiske review ikke har stærk evidens med hensyn til retning og styrke af effekter.

[6] Som en illustration heraf kan henvises til to forskningskortlægninger, som Clearinghouse har gennemført af skandinavisk institutionsforskning i årene 2006 og 2007. De registerer, at under 10 % af de i alt 107 studier, der foreligger, anvender et eksperimentelt design (Nordenbo et al., 2008; Nordenbo et al., 2009).

4.4.2 Metode ved syntesedannelse og evidensvægt

Ser vi derefter på den metode, der er anvendt ved syntesedannelsen, og den evidensvægt de enkelte grupperinger fremtræder med, giver det følgende resultat, jf. Tabel 4.5

Relation	Fag	Antal "high"	Antal "medium"	I alt	Relativ vægt
	Matematik	1	2	3	33 %
1	Naturfag	0	0	0	--
	Sprog	3	6	9	33 %
	Matematik	1	6	7	14 %
2	Naturfag	1	0	1	100 %
	Sprog	5	9	14	36 %
	Matematik	2	4	6	33 %
3	Naturfag	1	0	1	100 %
	Sprog	6	3	9	67 %
	Matematik	3	9	12	25 %
4	Naturfag	3	3	6	50 %
	Sprog	11	11	22	50 %
	Matematik	4	5	9	44 %
5	Naturfag	1	2	3	33 %
	Sprog	8	6	14	57 %

Tabel 4.5: Fordeling mellom evidensvekt "high" og "medium" i de forskellige narrative synteser

Ved gennemførelsen af en narrativ syntese er der først taget udgangspunkt i studier med tildelt evidensvægt "høj". Derefter er studier med evidensvægt "medium" – i den udstrækning, det har været muligt – blevet relateret hertil. Det fremgår af det ovenfor sagte, at antal studier og evidensvægt er højest for de narrative synteser, som berører relation 4 og 5, altså de studier hvor lærermeninger spiller en betydelig rolle i forbindelse med vurderingen af effekter på undervisning og elever af annonceringen af hændelsen 'at der skal gennemføres en test'. Et andet tydeligt træk er manglen på studier, som dokumenterer relation 1, og 3, altså den lineære kæde af processen fra testdata som kommunikeres til læreren (relation 1), noget som bidrager til at læreren responderer i form af en eller anden pædagogisk intervention (relation 2), som derefter påvirker elevernes læringsresultater og/eller har en anden virkning på eleven (relation 3). Når materialet splittes op i fag, er det tydeligt, at heller ikke de helt fagspecifikke virkemekanismer i den anvendte model er belyst af et stort antal undersøgelser. Sprog fremstår her som langt bedre dækket end både naturfag og matematik, men det bør erindres, at der bag denne fællesbetegnelse skjuler sig studier både inden for undervisning i modersmål (første sprog), fremmedsprog (andet og tredje sprog) og studier som specifikt er rettet mod henholdsvis læsning, skrivning og litteratur. De narrative synteser af relation 4 og 5 er derfor de mest robuste af dem, som er præsenteret i dette review, og man bør derfor være mere forsigtig med at drage konklusioner fra den del af reviewet, som omhandler den primære didaktiske akse i modellen (fra testdata, via lærers intervention og til elevlæring).

Et centralt aspekt ved reviewgruppens arbejde er, at gruppen bevidst blev sammensat for at repræsentere kompetencer inden for forskellige fag og relationer. Man kan derfor som udgangspunkt tænke sig, at de evidensvægte, som ligger bag hver enkelt af de narrative synteser, er påvirket af de enkelte reviewmedlemmers idiosynkratiske anvendelse af de forskellige vægte: høj, medium og lav. I tillæg vil der ved sådanne vurderinger altid eksistere et element af tilfældighed. Med andre ord er der en mulighed for, at evidensvægten indenfor hver enkelt af de narrative synteser ved vurderingen kan være påvirket af tilfældige og/eller systematiske fejlkilder. For at kunne dokumentere indslaget af tilfældighed ved vurderingen af evidensvægt, blev alle de inkluderede studier i tillæg bedømt af en af medarbejderne ved Clearinghouse. For yderligere at undgå systematiske fejlkilder, blev det i de tilfælde, hvor det var uenighed mellem de to bedømmelser, foretaget en konsensusbedømmelse. Da den samlede evidensvægten er baseret på tre mere specifikke vægte (se Tabel 3.17), blev både de specifikke og underordnede vægte og den samlede vægt bedømt uafhængigt af to personer for et bestemt udvalg af de inkluderede studiers vedkommende. Vi rapporterer dog her kun graden af uoverensstemmelse i den samlede evidensvægten, da det kun er den, der er blevet anvendt som inklusionskriterium i de narrative synteser. Tildelingen af lav, medium eller høj evidensvægt kan siges at repræsentere et underliggende kontinuum, men det er overgangen fra lav til medium, som er den kritiske. Studier som er blevet karakteriseret til at være af lav evidensvægt, vil jo i sidste instans ikke have nogen indvirkning på de narrative synteser og de konklusioner, som drages i reviewet.

I undersøgelsen af reviewet reliabilitet inkluderes derfor kun de studier, som et reviewmedlem har tildelt medium eller lav evidensvægt. Til sammen er dette 40 studier. For 7 af disse var der uoverensstemmelse ved, at en studie blev bedømt til at have henholdsvis lav eller medium evidensvægt. Det var enighed i vurderingen af de resterende 31 studier (83 %). Ganske vist kan man tænke sig, at man ved tilfældighed også får en lignende vurdering i nogle tilfælde, noget som ikke afdækkes i det enkle mål angivet som procentandel enighed. Cohen's kappa (Cohen, 1960) er en meget konservativ og omstridt indeks som korrigerer for dette, og i dette reliabilitetsstudie kan denne beregnes til 0,53. Da der ikke blev foretaget en tilsvarende vurdering mellem høj og medium evidensvægt, er det værd at bemærke, at ingen af studierne, som reviewmedlemmerne har tildelt medium evidensvægt, blev tildelt høj evidensvægt af den anden bedømmer. Dette tolkes som en indikation på, at grænsen mellem høj og medium evidensvægt er godt operationaliseret i reviewspørgsmålene og retningslinjerne i EPPI-revieweren. Det er heller ingen tendens til, at reviewmedlemmerne var systematisk strengere eller mildere i deres bedømmelse end medarbejderne ved Clearinghouse. Da uoverensstemmelserne derefter blev løst gennem konsensus, kan vi slå fast, at sandsynligheden for tilfældige fejl i den endelige vurdering er endnu lavere, end den ovenstående analysen antyder.

4.4.3 Undersøgelsens robusthed

Ser vi endelig på det tredje aspekt vedrørende undersøgelsens robusthed, drejer det sig om den grad af *information* om primærstudierne, der har tilladt inklusion i det systematiske review. I Kapitel 2 og 6, Appendiks 1, er der detaljeret redegjort for det foreliggende systematiske reviews begrebsmæssige afgrænsninger, søgning og anvendte søge-

profiler, screening og in/-eksklusions-principper, og for uddragningen af data fra de kortlagte studier.

4.4.4 Samlet vurdering

De her opregnede forhold tillader ikke nogen talmæssig beregning af de narrative syntesers robusthed udover de koefficienter, som er angivet for stabiliteten i bedømmelserne, jf. afsnit 4.4.2. Der må derfor foretages et skøn. Det er allerede på side 57 blev konstateret, at der i de sidste, små 30 år kun er gennemført få randomiserede, kontrollerede eksperimenter om det reviewspørgsmål, som denne undersøgelse vedrører, og at ingen af dem kan indgå i en meta-analyse. Derfor er syntesedannelsen sket som en narrativ analyse. Undersøgelsen af den gennemførte narrative analyses robusthed konkluderer, at undersøgelserne belyser relationerne 4 og 5 med større evidensvægt end relationerne 1, 2 og 3. Det blev samtidigt konstateret, at man bør være forsigtig med at drage konklusioner fra den del af reviewet, som omhandler den primære didaktiske aksen i modellen (fra testdata, via lærers intervention og til elevlæring)

Narrative synteser over de sidste små 30 års forskning på området tyder dog på, at der kan påvises mere almene træk og tendenser i denne forskning. Det er på den ene side påfaldende, at den gennemførte forskning giver et ret ensartet billede, der skal trækkes op i det efterfølgende Afsnit 4.5, et billede der samler sig om tre** hovedpunkter, suppleret med en række ekspliciteringer eller uddybninger.

Det må på den anden side ikke overses, at en forskningskortlægning af de sidste, små 30 års forskning på et givet område også kan tolkes som en afspejling af herskende faglige opfattelser eller forventninger hos forskere og opdragsgivere på området. Den kortlagte forskning giver ud fra denne betragtning et billede af, hvad forskere og opdragsgivere fandt værd at forske i, og hvilke rammer og svar de anså for frugtbare.

4.5 Afsluttende bemærkninger - om "pædagogisk brug af test"

En elev, der udsættes for test, påvirkes på en kompliceret måde både med hensyn til læringspraksis og læringsresultat. Den foreliggende undersøgelse er afgrænset til at se på de påvirkninger, som primært kommer

- fra annoncering af, at der skal testes, og
- fra de data, som testen genererer (testdata).

Bestræbelsen har været, at finde frem til de forhold der er relevante for lærerens brug af test i tilknytning til de enkelte elever og den klasse, læreren underviser i.

Til organisering af de undersøgelser, der i det systematiske review kan belyse indvirkning på elevlæringen, er anvendt en konceptuel model, der gengiver tre "veje", som påvirkningen på eleven kan løbe ad. Derved er undersøgelsen blevet splittet op i tre delundersøgelser, der henholdvis vedrører

- pædagogisk brug af testdata
- pædagogisk brug af test – virkninger på undervisningen, og
- pædagogisk brug af test – virkninger på eleven.

Det blev for det første konstateret, at ingen studier *på samme tid* fortæller os, hvilken viden lærere får fra testdata, hvordan de anvender dem til at udforme deres pædagogisk/didaktisk intervention, og hvordan den indlæringsmæssige effekt er af, at læreren har anvendt testdata pædagogisk.

Derimod er der påvist svag evidens for, at lærere foretrækker så uddybede oplysninger om testdata som muligt, eventuelt ledsaget af respons-/fejltype-analyser, og at lærere foretrækker test på klasseniveau, der er tilpasset den aktuelle undervisning, frem for formelle test. Der er ligeledes svag evidens for, at lærere anvender test til at planlægge og tilrettelægge undervisningen på klasseniveau. Dette er især tilfældet, hvis læreren ikke kender eleverne godt. Lærere anvender også test individuelt i forbindelse med karaktergivning, elev- og forældresamtaler, til belysning af faglige problemområder, og pædagogiske tiltag knyttet hertil.

Der er ligeledes svag evidens for, at det pædagogiske tiltag alene at rette og korrigere fejl i test ikke i sig selv medfører øget elevlæring. Der er også svag evidens for, at pædagogiske tiltag, hvor læreren ud fra testdata underviser i strategier, der letter indlæringsopgaven, giver en øget elevlæring.

Den samlede konklusion er, at den tolkning af "pædagogisk brug af test" som "pædagogisk brug af testdata" er svagt belyst forskningsmæssigt, idet en række ledsagende forhold dog er belyst med svag evidens.

Forskningsmæssigt er spørgsmålet om, hvordan annoncering af, at der skal testes, indvirker på undervisningen og eleven i sammenlignen med forrige problemstilling bedre belyst i den eksisterende forskning.

Undersøgelserne behandler wash-back effekten af at introducere formelle test. Overordnet viser de, at relationen mellem testning og undervisning er kompleks og i høj grad afhængig af kontekstuelle faktorer.

Derudover er en række forhold påvist med svag evidens, nemlig positive wash-back effekter – fra svagere til stærkere - med hensyn til testenes indflydelse på lærerens undervisning i intenderet retning, og betydeligere negative wash-back effekter af introduktion af formelle test. Disse effekter fordeler sig på tre områder: (a) indsnævret eller fordrejet curriculum, idet faglige tankegange forsimples, faktaviden og mekaniske færdigheder betones på bekostning af kreative og æstetiske perspektiver (b) undervisningstid allokeres til det/de fag, der skal testes i, på bekostning af de fag, der ikke testes i, og (c) undervisningen kan forfalde til træning til test og udenadslæren.

En samlet vurdering viser, at fordelene ved introduktion af formelle test kan opvejes af de negative virkninger, som disse test har især for de svagere præsterende elevers vedkommende.

Endelig er der svag evidens for, at elevernes testresultater stiger ved indførelse af test, men først efter nogle år. At annonceringen af, at der skal testes kan udløse følelsesreaktioner som nervøsitet og angst, at eleven forbereder sig på testen ved at lære udenad og memorere sætninger. For bedre præsterende elever stiger motivationen, mens svagere præsterende taber modet. Der er stærk evidens for, at de testdata, som testen genererer, kan virke ind på elevens fremtidige motivation og selvværd. Højt præsterende elever styrkes i motivation og selvværd, mens det modsatte er tilfældet for svagt præstende elever.

5 Konklusioner/ anbefalinger

5.1 Det systematiske reviews resultater

I dette systematiske review besvares følgende reviewspørgsmål:

> *Hvordan kan grundskolelæreres individ- og klassecentrerede brug af data fra test forbedre læreres didaktiske og/eller fagdidaktiske tiltag i klasser med almindelige elever? – spørgsmålet afgrænses til alene at inddrage testtyper, som indgår i de nationale test i de nordiske lande, og*
>
> *Hvordan indvirker indførelsen af testning på læreres didaktiske beslutninger og elevers læringsadfærd?*

Svaret er givet ved dels at gennemføre en forskningskortlægning og en narrativ syntese på baggrund af de sidste 30 års pædagogiske, empiriske forskning.

Svaret lyder,

- *at grundskolelæreres individ- og klassecentrerede brug af testdata i klasser med almindelige elever forbedres, når de didaktiske og/eller fagdidaktiske tiltag bygger på test, som lærerne oplever et medejerskab for.*
- *at introduktion af test fra instanser over klasseniveau med det formål at indvirke på undervisning og elevers læringsadfærd kan anvendes, idet der kan registeres både positive og negative virkninger på curriculum, undervisningen og tidsforbrug, på elevernes følelsesreaktioner, motivation, læringsadfærd og selvværd. Vurderet samlet opvejes de positive virkninger af central administerede test af de negative virkninger, især for de svagt præsterende elevers vedkommende.*

I rapportens afsnit 4.2.1, 4.2.2, 4.2.3 redegøres der i detaljer for dette svar. Til dette svar kan følgende bemærkninger om dets udsagnskraft knyttes:

- Svaret er baseret på den bedst tilgængelig evidens, der kan fremskaffes i den pædagogiske og uddannelsesvidenskabelige forskning i perioden 1980-2008
- Svaret bygger på en forskningskortlægning og forskningsvurdering af denne forskning
- Svaret er fremkommet ved at foretage narrative analyser på baggrund af de datauddragninger, som en reviewgruppe og Clearinghouse har foretaget.

Svaret giver anledning til følgende kommentarer:

- Svaret siger ikke noget om test anvendt som redskab til summativ evaluering, kvalitetskontrol eller benchmarking.
- Svaret siger om test anvendt som pædagogisk/didaktisk redskab i tilknytning til formativ evaluering, at de test, der er forankret i klassens daglige liv og lærernes pædagogiske tilrettelæggelse af undervisningen, har større chance for at blive anvendt pædagogisk end test, der er centralt administerede.

- Svaret sig om lærerne, at den information, som kan fås fra centralt administrerede og vedtagne test, kan bedres ved, at der sammen med testdata tilbydes respons- og/eller fejltypeanalyser af elevernes besvarelser.
- Svaret antyder, at læreruddannelsen bør inddrage lærerkompetencer til både at udvikle lokale test og til at tolke faglige analyser af testdata fra centralt producerede instanser.

5.2 Anbefalinger for praksis, policy og forskning

Til afslutning skal det overvejes, hvilke anbefalinger for praksis, policy og forskning, der udspringer af resultaterne fra det her gennemførte systematiske review.

5.2.1 Praksis

Undervisere skal være opmærksomme på, at centralt administrerede og pålagte test generer testdata, hvis information skal tolkes. Den eksisterende forskning siger ikke meget om, hvorledes det sker, eller hvordan den potentielle information, som rummes i testdata, eventuelt kan udnyttes.

Undervisere skal være opmærksomme på, at det pædagogiske tiltag blot at rette og korrigere fejl i test ikke i sig selv medfører øget elevlæring. Derimod tyder forskningen på, at pædagogiske tiltag, hvor læreren ud fra en analyse og diagnose af enkeltelevers testdata underviser i strategier, der letter elevens indlæringsopgave, derved bidrager til øget elevlæring.

Undervisere skal ligeledes være opmærksomme på, at centralt administrerede og pålagte test kan have både positive og negative virkninger. Især bør opmærksomheden rettes mod de negative virkninger på især de svagt præsterende elever i form af reduceret læringsadfærd, manglende motivation og selvværd. Foranstaltninger, der kan modvirke disse effekter, bør inddrages i den pædagogiske tilrettelæggelse og dagligdag.

Det kan derfor anbefales undervisere at forholde sig til en række af de konkrete forhold, som det systematiske review har påvist, kan være af betydning for deres elevers læring.

5.2.2 Policy

For beslutningstagere og tilrettelæggere af uddannelsespolitik, politikere og embedsmænd, kan det foreliggende systematiske reviews resultater hjælpe til udpege de styrker og svagheder, som anvendelsen af centralt administerede og pålagte test kan medføre som pædagogisk/didaktisk redskab.

Det bør samtidig understreges, at resultatet fra det systematiske review alene besvarer spørgsmålet om, hvilken pædagogisk brug af centralt administrerede og pålagte test forskningen har undersøgt. Det systematiske review har ikke beskæftiget sig med spørgsmålet om, hvorvidt test af denne type bør anvendes eller ej.

Det kan derfor anbefales, at politikere og embedsmænd, der ønsker at fremme brug af test som pædagogisk redskab til fremme af elevlæring, inddrager resultatet af det fore-

liggende systematiske review både med hensyn til styrker og svagheder ved sådanne test.

5.2.3 Forskning

Det foreliggende systematiske review har kortlagt de sidste 30 års empiriske forskning om pædagogisk brug af test.

Forskningskortlægning og forskningsvurdering har vist, at den pædagogiske og uddannelsesvidenskabelige forskning i disse år har demonstreret interesse for denne problemstilling, men at meget står tilbage at ønske. Det er flere gange blevet noteret, at der kun eksisterer få randomiserede, kontrollerede undersøgelser om begrænsede problemstillinger på dette område. Og at det i det hele kunne være ønskeligt, at der forelå undersøgelser af hovedproblemstillingen i dette systematiske review, nemlig den der vedrører den primære didaktiske akse i den konceptuelle model fra testdata, over lærerens intervention og til elevlæring. Dette systematiske review har godtgjort, at der p.t. ingen forskningsmæssig viden findes herom.

Desuden bør der gennemføres forskning, som anvender metoder, der sikrer en bedre indsigt i virkningen på undervisning og elever af gennemførelse af centralt administrerede og pålagte test. Den eksisterende forskning bygger i for høj grad på forskningsdesign, der trækker på informanternes vurdering af disse virkninger. Som det er godtgjort af den primærforskning, der indgår i dette systematiske review, kan der være diskripans mellem informantens mening om egen adfærd og observationen af den faktiske adfærd.

Endelig peger resultaterne fra det systematiske review på, at der bør gennemføres primærforskning til belysning af alle de tre "veje", som kan indvirke på elevens læring. I denne forskning bør også inddrages registrering af testtypernes betydning for denne indvirkning.

Det kan derfor anbefales at iværksætte empirisk primærforskning om pædagogisk brug både i betydningen brug af testdata og virkningerne på undervisning og elever af introduktion af test under anvendelse af forskningsdesign, der med størst evidens kan redegøre for disse sammenhænge.

6 Appendiks 1: Søgeprofiler

I afsnit 2.3 er de enkelte databaser og ressourcer nøjere beskrevet. Herunder gengives alene de anvendte søgeprofiler.

1. ERIC: ERIC1 eller ERIC2 eller ERIC3 eller ERIC4 eller ERIC5 eller ERIC6

ERIC1 (tests i matematik):

Query: (DE=("mathematics" or "algebra" or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or "modular arithmetic" or "multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT =(143 reports: research)))

ERIC2 (tests i læsning - ikke i matematik)

Query: ((DE=("reading" or "basal reading" or "beginning reading" or "content area reading" or "corrective reading" or "critical reading" or "directed reading activity" or "early reading" or "functional reading" or "independent reading" or "individualized reading" or "music reading" or "oral reading" or "reading fluency" or "reading aloud to others" or "recreational reading" or "remedial reading" or "silent reading" or "speed reading" or "story reading" or "sustained silent reading" or "reading ability" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading achievement" or "reading assignments" or "reading attitudes" or "reading comprehension" or "reading diagnosis" or "miscue analysis" or "reading difficulties" or "reading habits" or "reading improvement" or "reading instruction" or "basal reading" or "content area reading" or "corrective reading" or "directed reading activity" or "individualized reading" or "remedial reading" or "sustained silent reading" or "reading materials" or "large type materials" or "supplementary reading materials" or "telegraphic materials" or "reading motivation" or "reading processes" or "decoding reading" or "reading programs" or "adult reading programs" or "reading rate" or "reading readiness" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading strategies" or "reading teachers")) AND ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national compe-

tency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT ((DE=("mathematics" or "algebra" or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or "modular arithmetic" or "multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT =(143 reports: research))))

ERIC3 (tests i naturvidenskaber ikke i matematik eller læsning)

Query: (((DE=("physics" or "biophysics" or "biomechanics" or "bionics" or "robotics" or "electronics" or "microelectronics" or "mechanics physics" or "fluid mechanics" or "kinetics" or "diffusion physics" or "quantum mechanics" or "nuclear physics" or "optics" or "thermodynamics" or "chemistry" or "biochemistry" or "geochemistry" or "inorganic chemistry" or "organic chemistry" or "physical chemistry" or "electrochemistry" or "stereochemistry")) or (DE=("physical sciences" or "natural sciences"))) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT (((DE=("reading" or "basal reading" or "beginning reading" or "content area reading" or "corrective reading" or "critical reading" or "directed reading activity" or "early reading" or "functional reading" or "independent reading" or "individualized reading" or "music reading" or "oral reading" or "reading fluency" or "reading aloud to others" or "recreational reading" or "remedial reading" or "silent reading" or "speed reading" or "story reading" or "sustained silent reading" or "reading ability" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading achievement" or "reading assignments" or "reading attitudes" or "reading comprehension" or "reading diagnosis" or "miscue analysis" or "reading difficulties" or "reading habits" or "reading improvement" or "reading instruction" or "basal reading" or "content area reading" or "corrective reading" or "directed reading activity" or "individualized reading" or "remedial reading" or "sustained silent reading" or "reading materials" or "large type materials" or "supplementary reading materials" or "telegraphic materials" or "reading motivation" or "reading processes" or "decoding reading" or "reading programs" or "adult reading programs" or "reading rate" or "reading readiness" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading strategies" or "reading teachers")) AND ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT ((DE=("mathematics" or "algebra" or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or

"modular arithmetic" or "multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))))

ERIC4: (matematiktests eller læsetests eller naturvidenskabstests. Ikke: tests i matematik, læsning eller naturvidenskab)

Query: ((PT=(142 reports: evaluative) or PT=(143 reports: research)) and (DE=("mathematics tests" or "reading tests" or "science tests"))) NOT (((((DE=("physics" or "biophysics" or "biomechanics" or "bionics" or "robotics" or "electronics" or "microelectronics" or "mechanics physics"or "fluid mechanics" or "kinetics" or "diffusion physics" or "quantum mechanics" or "nuclear physics" or "optics" or "thermodynamics" or "chemistry" or "biochemistry" or "geochemistry" or "inorganic chemistry" or "organic chemistry" or "physical chemistry" or "electrochemistry" or "stereochemistry")) or (DE=("physical sciences" or "natural sciences"))) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT (((DE=("reading" or "basal reading" or "beginning reading" or "content area reading" or "corrective reading" or "critical reading" or "directed reading activity" or "early reading" or "functional reading" or "independent reading" or "individualized reading" or "music reading" or "oral reading" or "reading fluency" or "reading aloud to others" or "recreational reading" or "remedial reading" or "silent reading" or "speed reading" or "story reading" or "sustained silent reading" or "reading ability" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading achievement" or "reading assignments" or "reading attitudes" or "reading comprehension" or "reading diagnosis" or "miscue analysis" or "reading difficulties" or "reading habits" or "reading improvement" or "reading instruction" or "basal reading" or "content area reading" or "corrective reading" or "directed reading activity" or "individualized reading" or "remedial reading" or "sustained silent reading" or "reading materials" or "large type materials" or "supplementary reading materials" or "telegraphic materials" or "reading motivation" or "reading processes" or "decoding reading" or "reading programs" or "adult reading programs" or "reading rate" or "reading readiness" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading strategies" or "reading teachers")) AND ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT ((DE=("mathematics" or "algebra" or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or "modular arithmetic" or

"multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))))) or (((DE=("reading" or "basal reading" or "beginning reading" or "content area reading" or "corrective reading" or "critical reading" or "directed reading activity" or "early reading" or "functional reading" or "independent reading" or "individualized reading" or "music reading" or "oral reading" or "reading fluency" or "reading aloud to others" or "recreational reading" or "remedial reading" or "silent reading" or "speed reading" or "story reading" or "sustained silent reading" or "reading ability" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading achievement" or "reading assignments" or "reading attitudes" or "reading comprehension" or "reading diagnosis" or "miscue analysis" or "reading difficulties" or "reading habits" or "reading improvement" or "reading instruction" or "basal reading" or "content area reading" or "corrective reading" or "directed reading activity" or "individualized reading" or "remedial reading" or "sustained silent reading" or "reading materials" or "large type materials" or "supplementary reading materials" or "telegraphic materials" or "reading motivation" or "reading processes" or "decoding reading" or "reading programs" or "adult reading programs" or "reading rate" or "reading readiness" or "reading skills" or "reading comprehension" or "reading fluency" or "reading rate" or "reading strategies" or "reading teachers")) AND ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))) NOT ((DE=("mathematics" or "algebra" or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or "modular arithmetic" or "multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research))))) or ((DE=("mathematics" or "algebra"

or "matrices" or "polynomials" or "vectors mathematics" or "arithmetic" or "addition" or "division" or "modular arithmetic" or "multiplication" or "subtraction" or "calculus" or "ethnomathematics" or "geometry" or "analytic geometry" or "plane geometry" or "polygons" or "triangles geometry" or "solid geometry" or "topology" or "probability" or "statistics" or "bayesian statistics" or "least squares statistics" or "maximum likelihood statistics" or "nonparametric statistics" or "sampling" or "item sampling" or "statistical distributions" or "technical mathematics" or "trigonometry" or "mathematics achievement" or "mathematics activities" or "mathematics curriculum" or "college mathematics" or "elementary school mathematics" or "general mathematics" or "numeracy" or "modern mathematics" or "secondary school mathematics" or "mathematics education" or "mathematics instruction" or "remedial mathematics" or "mathematics skills" or "numbers" or "logarithms" or "number systems" or "exponents mathematics" or "rational numbers" or "fractions" or "decimal fractions" or "integers" or "prime numbers" or "whole numbers" or "reciprocals mathematics")) and ((DE=("achievement tests" or "equivalency tests" or "mastery tests" or "national competency tests")) and (PT=(142 reports: evaluative) or PT=(143 reports: research)))))

ERIC5:

washback* or backwash*

ERIC6:

consequential validity

2. BEI: BEI1 eller BEI2 eller BEI3

BEI1

Q-P-PY=("2000" OR "2001" OR "2002" OR "2003" OR "2004" OR "2005" OR "2006" OR "2007") OR (Q-P-PY=("1980" OR "1981" OR "1982" OR "1983" OR "1984" OR "1985" OR "1986" OR "1987" OR "1988" OR "1989" OR "1990" OR "1991" OR "1992" OR "1993" OR "1994" OR "1995" OR "1996" OR "1997" OR "1998" OR "1999")) AND (Q-W-00=MATH? OR ARITMETIC? OR SCIENCE? OR PHYSICS OR CHEMISTRY OR READING AND (Q-P-ZZ=("NATIONAL COMPETENCY TESTS") OR (Q-P-00=("EQUIVALENCY TESTS" OR "MASTERY TESTS")) OR (Q-P-ZZ=("ACHIEVEMENT TESTS"))))

BEI2

Q-P-PY=("2000" OR "2001" OR "2002" OR "2003" OR "2004" OR "2005" OR "2006" OR "2007") OR (Q-P-PY=("1980" OR "1981" OR "1982" OR "1983" OR "1984" OR "1985" OR "1986" OR "1987" OR "1988" OR "1989" OR "1990" OR "1991" OR "1992" OR "1993" OR "1994" OR "1995" OR "1996" OR "1997" OR "1998" OR "1999")) AND (Q-P-ZZ=("READING TESTS") OR (Q-P-ZZ=("MATHEMATICS TESTS")) OR (Q-P-ZZ=("SCIENCE TESTS")))

BEI3

WASHBACK? OR BACKWASH? OR CONSEQUENTIAL VALIDITY

3. AEI: AEI1 eller AEI2 eller AEI3

AEI1

Q-P-PY=("1999" OR "1999?" OR "2000" OR "2001" OR "2001?" OR "2002" OR "2002?" OR "2003" OR "2004" OR "2005" OR "2006" OR "2007") OR (Q-P-PY=("1980" OR "1981" OR "1982" OR "1983" OR "1984" OR "1985" OR "1986" OR "1987" OR "1988" OR "1989" OR "1990" OR "1991" OR "1992" OR "1993" OR "1994" OR "1995" OR "1996" OR "1997" OR "1998" OR "1998?")) AND (Q-W-00=MATH? OR ARITHMETIC? OR SCIENCE? OR PSYSICS OR CHEMISTRY OR READING AND (Q-P-ZZ=("EQUIVALENCY TESTS") OR (Q-P-00=("MASTERY TESTS" OR "NATIONAL COMPETENCY TESTS")) OR (Q-P-ZZ=("ACHIEVEMENT TESTS"))))

AEI2

Q-P-PY=("1999" OR "1999?" OR "2000" OR "2001" OR "2001?" OR "2002" OR "2002?" OR "2003" OR "2004" OR "2005" OR "2006" OR "2007") OR (Q-P-PY=("1980" OR "1981" OR "1982" OR "1983" OR "1984" OR "1985" OR "1986" OR "1987" OR "1988" OR "1989" OR "1990" OR "1991" OR "1992" OR "1993" OR "1994" OR "1995" OR "1996" OR "1997" OR "1998" OR "1998?")) AND (Q-P-ZZ=("SCIENCE TESTS") OR (Q-P-00=("READING TESTS")) OR (Q-P-00=("MATHEMATICS TESTS")))

AEI3

Q-W-00=BACKWASH? OR WASHBACK? OR CONSEQUENTIAL VALIDITY AND (Q-P-PY=("1999" OR "1999?" OR "2000" OR "2001" OR "2001?" OR "2002" OR "2002?" OR "2003" OR "2004" OR "2005" OR "2006" OR "2007") OR (Q-P-PY=("1980" OR "1981" OR "1982" OR "1983" OR "1984" OR "1985" OR "1986" OR "1987" OR "1988" OR "1989" OR "1990" OR "1991" OR "1992" OR "1993" OR "1994" OR "1995" OR "1996" OR "1997" OR "1998" OR "1998?")))

4. CBCA-education: CBCA1 eller CBCA2 eller CBCA3

CBCA1

(achievement tests OR competency tests) AND (math* OR arithmetic* OR science* OR chemistry OR physics OR reading)

CBCA2

matematics test*" OR "reading test*" OR "science test*

CBCA3

backwash* OR washback* OR consequential validity

5. Psychinfo: Psychinfo1 eller Psychinfo2

Psychinfo1

(DE=("mathematics education" or "science education" or "reading education" or "physics" or "chemistry")) and (DE=("achievement measures" or "iowa tests of basic skills" or "stanford achievement test" or "wide range achievement test" or "woodcock johnson psychoeducational battery"))

Psychinfo2

washback or backwash* or "consequential validity”

6. FIS Bildung: FIS Bildung1 eller FIS Bildung2 eller FIS Bildung3

FIS Bildung1

(mathemati* lesen* physik* chemie* naturwissenschaft* /ODER) UND (schulleistungsmessung *leistungstest*/ODER)

FIS Bildung2

mathematiktest* rechnentest* lesetest* naturwissenschafttest* physiktest* chemie-test/ODER

FIS Bildung3

"konsequenz* validität" backwash* washback* /ODER

7. JYKDOK: JYKDOK1 eller JYKDOK2 eller JYKDOK3 eller JYKDOK4 eller JYKDOK5

Profile 1, 2 limited to: English language material published i Finland 1980-2007

JYKDOK1

test?

AND

(math? OR arithmetic? OR science? OR physics OR chemistry OR reading)

JYKDOK2

washback? OR backwash? OR "consequen? validity"

Profile 3, 4,and 5 limited to: Swedish language material published in Finland 1980-2007

JYKDOK3

(kunskapstest? OR test? OR prov? OR kompetenstest?)

AND

(matemati? OR aritmeti? OR fysik? OR kemi? OR läs?)

JYKDOK4

matematiktest? OR aritmetiktest? OR naturvetenskapstest? OR kemitest? OR fysiktest? OR lästest? OR matematikprov? OR aritmetikprov? OR naturvetenskapsprov? OR kemiprov? OR fysikprov? OR läsprov?

JYKDOK5

washback? OR backwash? OR konsekvensvalid? OR "konsekvens validitet?"

8. Evidensbasen:

test? or assessment

9. News Alerts DPB:

Indholdsfortegnelser af nye tidsskrifter er gennemset og relevante artikler uploaded i EPPI Reviewer.

10. Libris: Libris1 eller Libris2 eller Libris3

Libris beta version search:

Only material published 1980- included

Libris1

”land:sw tree:e (kunskapstest* OR test* OR prov* OR kompetenstest*) NOT testamente"

Libris2

”land:sw (matematiktest* OR matematikprov* OR Aritmetiktest* OR aritmetikprov* OR Naturvetenskapstest* OR naturvetenskapsprov* OR kemitest* OR kemiprov* OR fysiktest* OR fysikprov* OR lästest* OR läsprov*)"

Libris3

”land:sw (washback* OR backwash* OR konsekvensvalid* OR “konsekvens valid*”)"

11. Dansk Pædagogisk Base: Profile 1 eller profile 2 eller profile 3

Profile 1

kd=pep? og (test? eller prøve? eller standpunktsprøve? eller kompetencetest? eller kompetenceprøve?) og (dk=37.1? eller dk=37.3?)

Profile 2

Kd=pep? og (matematiktest? eller matematikprøve? eller regneprøve? eller regnetest? eller naturvidenskabstest? eller naturvidenskabsprøve? eller kemitest? eller kemiprøve? eller fysiktest? eller fysikprøve? eller læsetest? eller læseprøve?)

Profile 3

Kd=pep? og (washback? OR backwash? OR konsekvensvalid? OR “konsekvens valid?”)

12. NORBOK: Profile 1 eller profile 2 eller profile 3

Profile 1

(kunnskapstest? OR kunnskapsprøve? OR prøve?” OR test? OR ferdighetsprøve? OR kompetansetest?) AND Dewey classification:37? NOT testamente?

Profile 2

matematikktest? OR regnetest? OR naturvitenskapstest? OR kjemitest? OR fysikktest? OR lesetest? OR matematikkprøve? OR regneprøve? OR naturvitenskapsprøve? OR kjemiprøve? OR fysikkprøve? OR leseprøve?

Giver kun 1 hit

Profile 3

washback? OR backwash? OR konsekvensvalid? OR “konsekvens valid?”

13. References from review Group:

Manuelt upload i EPPI Reviewer af alle forslag fra Reviewgruppe, som ikke allerede er uploaded.

14. Opdateringssøgninger:

er alle gennemført med samme profil som ovenstående grundsøgninger, men med begrænsning, der sikrer, at kun tilvæksten efter grundsøgninger findes.

7 Appendiks 2: Et eksempel på en genbeskrivelse

Item: Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. Applied Measure-ment in Education, 16(1), 1-26.

7.1 EPPI-Centre data extraction and coding tool for education studies V2.0

Section A: Administrative details

A.1 Name of the reviewer	Details *Jens Dolin, Neriman Tiftikci*
A.2 Date of the review	Details *Teh 25th of February 2008*
A.3 Please enter the details of each paper which reports on this item/study and which is used to complete this data extraction.	Paper (1) *Journal article* Unique Identifier: *Item number 780282* Authors: *Clement A. Stone, Suzanne Lane* Title: *Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables*
A.4 Main paper. Please classify one of the above papers as the 'main' report of the study and enter its unique identifier here.	Unique Identifier: *Item number 780282*
A.5 Please enter the details of each paper which reports on this study but is NOT being used to complete this data extraction.	
A.6 If the study has a broad focus and this data extraction focuses on just one component of the study, please specify this here.	Not applicable (whole study is focus of data extraction)
A.7 Identification of report (or reports)	Electronic database
A.8 Status	Published
A.9 Language (please specify)	Details of Language of report

	English

Section B: Study Aims and Rationale

B.1 What are the broad aims of the study?	Explicitly stated (please specify) *The purpose of this study was to explore the relationship between (a) changes in the scores from the Maryland State Performance Assessment Program from 1993 to 1998 and (b) classroom instruction and assessment practices, student learning and motivation, students' and teachers' beliefs about and attitudes toward the assessment, and a school characteristic. Several factors from each of these dimensions were found to explain a significant amount of variability in school performance over time using growth models.*
B.2 What is the purpose of the study?	B: Exploration of relationships
B.3 If the study addresses a 'what works' question, does it focus on outcomes or process?	
B.4 Why was the study done at that point in time, in those contexts and with those people or institutions?	Explicitly stated (please specify) *The nature of assessment programs has also changed with time to mirror the philosophies of different educational reform movements. In accordance with the most recent educational reform movement, a number of states are now implementing statewide assessment programs that involve performance-based tasks, in response to arguments that assessments utilizing more traditional types of standardized tests have led to educational practices that overemphasize basic skills (e.g., Resnick & Resnick, 1992). The prevailing assumption underlying the use of performance-based assessments is that they encourage the use of instructional strategies and techniques that foster reasoning, problem solving, and communication (National Council on Education Standards and Testing, 1992). One state implementing such an assessment program is Maryland.* *The Maryland State Performance Assessment Program (MSPAP) is designed to provide school-level information at Grades 3, 5, and 8 and to provide information for school accountability and improvement (Maryland State Board of Educa-*

	tion, 1995). Implemented in the early 1990s, the MSPAP requires students to develop written responses to interdisciplinary tasks that require the application of skills and knowledge to reallife problems, and it is intended to promote perform-ance-based instruction and classroom assessments. Given the high expectations for performance-based assessments, the consequences of the uses and interpretations of the assessments need to be addressed, including both (a) negative and positive conse-quences and (b) intended and plausible unintended consequences
B.5 Was the study informed by, or linked to, an existing body of empirical and/or theoretical re-search?	Explicitly stated (please specify) *Cronbach, 1988; Koretz, Barron, Mitchell, & Stecher, 1996; Linn, Baker, & Dunbar, 1991; Mes-sick, 1989; Stevens, 1999; Stevens, Estrada, & Parkes, 200 (Bryk & Raudenbush, 1992).*
B.6 Which of the following groups were consulted in working out the aims of the study, or issues to be addressed in the study?	None/Not stated Coding is based on: Reviewers' inference
B.7 Do authors report how the study was funded?	Not stated/unclear (please specify)
B.8 When was the study carried out?	Not stated/unclear (please specify)
B.9 What are the study research questions and/or hypotheses?	Implicit (please specify) *The purpose of this study was to examine this issue with regard to the MSPAP and to explore the relation-ship between (a) changes in MSPAP test scores for schools and (b) classroom instruction and assessment practices, student learning and motivation, students' and teachers' beliefs about and attitudes toward the MSPAP, and a school characteristic. Greater external validity was imparted to interpretations by analyzing data reflecting different MSPAP subject areas (mathematics, reading, writing, science, and social studies) for different sets of schools.*

Section C: Study Policy or Practice Focus

C.1 What is/are the topic focus/foci of the study?	Assessment (please specify) *The study assess the students's and teachers' beliefs about and attitudes toward the MSPAP* Teaching and learning (please specify) *investigation of classroom instruction and assessment practices, student learning and motivation, students' and teachers' beliefs about and attitudes toward the MSPAP, and a school characteristic.* Coding is based on: Reviewers' inference
C.2 What is the curriculum area, if any?	Literacy - first languages *reading and writing* Maths Science Coding is based on: Authors' description
C.3 What is/are the educational setting(s) of the study?	Primary school
C.4 In which country or countries was the study carried out?	Explicitly stated (please specify) *maryland, USA*
C.5 Please describe in more detail the specific phenomena, factors, services or interventions with which the study is concerned.	Details *Most states are implementing statewide assessment programs that are being used for high-stakes purposes. Some of these assessments involve performance-based tasks that are assumed not only to serve as motivators for improving student achievement and learning, but also to encourage instructional strategies and techniques in the classroom that are more consistent with reform-oriented educational outcomes. Given these high expectations, more comprehensive and direct evidence for the consequences of the assessments needs to be gathered*

Section D: Actual sample

D.1 Who or what is/ are the sample in the study?	Learners Teaching staff
D.2 What was the total number of participants in the study (the actual sample)?	Explicitly stated (please specify)

	School sample for mathematics and language arts (reading and writing). *Seventy-two elementary schools and 36 middle schools were selected to participate in the study, with alternate schools identified as potential replacements for schools whose administrators declined the offer to participate.* ***School sample for science and social studies***. *One hundred twentysix elementary schools and 63 middle schools were selected to participate in the study, with alternate schools identified as potential replacements for schools whose administrators declined the offer to participate.* ***Final sample for the study***. *As a result of questionnaire nonresponse, the linking of the questionnaire data to the mean scale scores for schools reduced the sample sizes. The final samples for this study consisted of 86 of 90 schools for mathematics and language arts (reading and writing), 116 of 161 schools for science, and 111 of 161 schools for social studies. The samples for science and social studies were reduced disproportionately more than for the other content areas because both teacher and student questionnaire data were linked to science and social studies MSPAP performance data (listwise deletion). Less reduction in the sample would have occurred if only teacher questionnaire data were analyzed with the science and social studies performance data. However, student evidential variables could not have been examined.*
D.3 What is the proportion of those selected for the study who actually participated in the study?	Explicitly stated (please specify) ***School sample for mathematics and language arts (reading and writing)*** *The final sample consisted of 59 elementary schools (82% participation) and 31 middle schools (86% participation), for a total of 90 schools. Of the 59 elementary schools, 42 were from the initial 72 that were sampled, and of the 31 middle schools, 22 were from the initial 36 that were sampled. The remaining schools were from the list of alternate schools for each classification cell. This sample represented schools from 19 of 24 systems or counties in Maryland, and there were approximately*

	equal numbers of schools (~8% to 12%) within each of the nine classifications cells. ***School sample for science and social studies*** *The final sample consisted of 103 elementary schools (82 % participation) and 58 middle schools (92% participation), for a total of 161 schools. Of the 103 elementary schools, 87 were from the initial 126 that were sampled, and of the 58 middle schools, 44 were from the initial 63 that were sampled. The remaining schools were from the list of alternate schools for each classification cell. This sample represented schools from 22 of 24 systems or counties in Maryland, and there were approximately equal numbers of schools (~9% to 13%) within each of the nine classification cells. In summary, during the 2 years, 251 schools participated in the study.* ***Final sample for the study.*** *As a result of questionnaire nonresponse, the linking of the questionnaire data to the mean scale scores for schools reduced the sample sizes. The final samples for this study consisted of 86 of 90 schools for mathematics and language arts (reading and writing), 116 of 161 schools for science, and 111 of 161 schools for social studies. The samples for science and social studies were reduced disproportionately more than for the other content areas because both teacher and student questionnaire data were linked to science and social studies MSPAP performance data (listwise deletion). Less reduction in the sample would have occurred if only teacher questionnaire data were analyzed with the science and social studies performance data. However, student evidential variables could not have been examined.*
D.4 Which country/countries are the individuals in the actual sample from?	Implicit (please specify) *USA*
D.5 If the individuals in the actual sample are involved with an educational institution, what type of institution is it?	Primary school (please specify) *elementary and middle school* Secondary school (please specify age range) Coding is based on: Authors' description

D.6 What ages are covered by the actual sample?	5-10 11-16 Coding is based on: Reviewers' inference
D.7 What is the sex of participants?	Mixed sex (please specify)
D.8 What is the socio-economic status of the individuals within the actual sample?	Implicit (please specify) *The study uses a "free" or "reduced lunch variable" as a proxy for socioeconomic status in the stratified random sampling procedure in order to classify shcools into one of three categories (lower third, middle third, and upper third). As such the socio-economic status of the individuals within the actual sample varies from low to high.*
D.9 What is the ethnicity of the individuals within the actual sample?	Not stated/unclear (please specify)
D.10 What is known about the special educational needs of individuals within the actual sample?	Not stated/unclear (please specify) *No special educational needs are reported.*
D.11 Please specify any other useful information about the study participants.	Details *Means of MSPAP scale scores for schools (linked to the 1993 assessment) in the sample from 1993 to 1997 or 1998 were provided by personnel within the MSDE.* *In this study, changes in MSPAP scale scores for schools were examined in relation to classroom instruction and assessment practices, student learning and motivation, beliefs about the impact of and attitudes toward the MSPAP, and the school*

Section E: Programme or Intervention description

E.1 If a programme or intervention is being studied, does it have a formal name?	Yes (please specify) *Maryland State Performance Assessment Program from 1993 to 1998 is being studied-*
E.2 Theory of change	Details *not an intervention*
E.3 Aim(s) of the intervention	Not stated *not an intervention*
E.4 Year intervention started	Details

	not an intervention
E.5 Duration of the intervention	Not stated *not an intervention*
E.6 Person providing the intervention (tick as many as appropriate)	Not stated *not an intervention*
E.7 Number of people recruited to provide the intervention (and comparison condition) (e.g. teachers or health professionals)	Not stated *not an intervention*
E.8 How were the people providing the intervention recruited? (Write in) Also, give information on the providers involved in the comparison group(s), as appropriate.	Not stated *not an intervention*
E.9 Was special training given to people providing the intervention?	Not stated *not an intervention*

Section F: Results and conclusions

F.1 How are the results of the study presented?	Details *chi square and tabels*
F.2 What are the results of the study as reported by the authors?	Details *Table 1 summarizes the mean MSPAP performance across schools (elementary and middle schools combined) in the sample. As can be seen, the general trend indicated one of increasing mean performance with time. Except in the case of writing, there appeared to be larger gains in the early years, followed by a leveling of the scores during 1995-1996, at which point there was again an increase in mean performance for schools. In addition, a decrease in mean performance was noted for science and social studies from the 1995 administration to the 1996 administration of the MSPAP.* *TABLE 1 Means and Standard Deviations of MSPAP Scale Scores Across Schools (1993-1998)* *Math Writing Reading Science Social tudies Adm. M SD M SD M SD M SD M SD* *1993 510.6 24.1 504.6 19.3 503.8 20.6 509.6 25.3 503.3 18.5* *1994 513.9 23.0 503.8 18.8 505.6 19.5 514.6 23.0 508.4 18.7*

	1995 518.1 22.8 506.1 20.1 512.4 17.1 519.0 21.4 512.8 17.0 *1996 518.7 22.9 511.3 20.0 513.0 17.6 518.3 23.8 511.5 19.0* *1997 521.8 24.1 513.9 22.5 517.1 17.5 518.9 24.9 514.4 20.0* *1998 523.6 22.9 518.7 19.4* *Note. The sample sizes (number of schools across grades) for the subject areas were as follows: Math (86), Writing (86), Reading (86), Science (116), and Social Studies (111). MSPAP = Maryland State Performance Assessment Program; Adm. = administration year.* *In 1993, schools in the lower quartile (which reflects higher SES) were concentrated in the range of* *520 to 550, whereas schools in the upper quartile (which reflects lower SES) were* *concentrated in the range of 480 to 500. In addition, the rate of change for schools* *in the lower quartile exhibited a more consistent increase with time, whereas considerably* *more variability was observed for schools in the upper quartile. In both* *cases, the rate of change appears modest from 1993 to 1998, a finding that is consistent* *with results for other assessment programs.* *The finding that students' perceptions of the degree to which they worked on* *MSPAP-like tasks was negatively related to 1998 MSPAP performance is interesting,* *and in the case of MSPAP science performance, there is an apparent paradox* *between the direction of the relationship for Current Instruction (b = 5.9) and* *students' perception of MSPAP-like Instruction (b = -4.0). As teachers' instruction* *more closely reflected the Maryland Learning Outcomes and reform-oriented* *problem types, higher 1998 MSPAP performance was observed. In contrast, student* *perception of the degree to which they worked on MSPAP-like tasks was* *negatively related to 1998 MSPAP performance. Several possible explanations* *exist for this finding. For example, students were administered the student questionnaire* *just after the MSPAP, and other studies indicated that MSPAP test preparation* *activities were more aligned with the MSPAP than instructional activities* *were (e.g., Cerrillo, Hansen, Parke, Lane, & Scott,*

	2000). Given the question "How often did you work on tasks like those on the MSPAP?" students may have been focusing on the recent MSPAP test preparation activities. Thus, perhaps lower performing schools use more MSPAP-like formatted tasks in test preparation activities than those used by higher performing schools. Another possible explanation is that given the question "How often did you work on tasks like those on the MSPAP?" students may have been focusing on the format of MSPAP tasks and not on the learning outcomes reflected in the tasks. If so, lower performing schools may have been more likely to use more MSPAP-like formatted tasks than those used by higher performing schools. Schools performing at higher levels may be more successful at reflecting the science learning outcomes in a variety of reform-oriented problem formats. The finding of a similar negative effect for social studies substantiates the direction of the MSPAP-like Instruction effect. Finally, teacher perceptions of the degree of MSPAP Impact on classroom reading and writing activities was found to explain significantly the variability in 1997 reading and writing performance. The direction of the effect indicates that an increased perceived impact of the MSPAP was associated with increased levels in 1997 MSPAP reading and writing performance.
F.3 What do the author(s) conclude about the findings of the study?	Details *Teacher reports for some instruction-related predictors (Current Instruction or Reform-Oriented Tasks) were found to explain differences in MSPAP performance levels in all subject areas except social studies. In addition, the Reform-Oriented Tasks dimension was found to explain differences in rates of change in MSPAP performance with time for reading and writing. Finally, the perceived general impact of the MSPAP on instruction and assessment practices was found to explain either differences in MSPAP performance levels or rates of change with time across all subject areas except Social Studies. On the basis of the same set of questionnaires, other analyses have also suggested that classroom instruction and assessment practices have changed with time with*

	the educational reform movement in Maryland (Lane et al., 2000). However, studies indicate that there are still gains to be made in the congruence between (a) classroom instruction and assessment practices and (b) the state-defined learning outcomes and the MSPAP (e.g., Cerrillo et al., 2000; Lane, Parke, & Stone, 1998). Student reports regarding use of MSPAP-like tasks. *The results of this study provide some correlational evidence for the positive impact of a statewide assessment program. Changes in scores from the MSPAP from 1993 to 1998 were found to be related to school, classroom, and student factors.*

Section G: Study Method

G.1 Study Timing	Retrospective *from 1993 to 1998 is being studied.*
G.2 when were the measurements of the variable(s) used as outcome measures made, in relation to the intervention	Only after
G.3 What is the method used in the study?	Cohort study Cross-sectional study Views study

Section H: Methods-groups

H.1 If Comparisons are being made between two or more groups*, please specify the basis of any divisions made for making these comparisons	No prospective allocation but use of pre-existing differences to create comparison groups
H.2 How do the groups differ?	Explicitly stated (please specify) *The data relevant to this study were obtained from questionnaires that were developed for teachers and students. Questionnaires specific to the different subject areas were developed for* ***elementary school (third- and fifth-grade) and middle school (eighth-grade)*** *teachers and students.* *A stratified random sampling procedure was used to select schools for the study.*

	The strata were defined by three levels for each of two variables: *(a) the percentage of free or reduced-price lunches according to the 1994–1995 classification used by the Maryland State Department of Education (MSDE) and* *(b) the MSPAP performance gains (the MSDE's 1993–1995 change index).* *Teachers completed questionnaires prior to administering the MSPAP, whereas students completed questionnaires within 2 weeks following the administration of the MSPAP.*
H.3 Number of groups	Two
H.4 If prospective allocation into more than one group, what was the unit of allocation?	Not applicable (no prospective allocation)
H.5 If prospective allocation into more than one group, which method was used to generate the allocation sequence?	Not applicable (no prospective allocation)
H.6 If prospective allocation into more than one group, was the allocation sequence concealed?	Not applicable (no prospective allocation)
H.7 Study design summary	Details ***Elementary and middle schools*** *are compared with each other. The divisions are made across following subjects* ***mathematics, language arts (reading and writing), science and social studies****. Furthermore the following variables are also considered: In this study, changes in MSPAP scale scores for schools were examined in relation to classroom instruction and assessment practices, student learning and motivation, beliefs about the impact of and attitudes toward the MSPAP, and the school characteristic of* ***percentage of free or reduced-price lunches****, which served as a proxy for SES. The MSDE also provided school data on the percentage of free or reduced-price lunches.*

Section I: Methods - Sampling strategy

I.1 Are the authors trying to produce findings that are representative of a given population?	Implicit (please specify) *Is not explicitly mentioned though this study is trying to build on the existing research about assessments:*

	Given the high expectations for performance-based assessments, the consequences of the uses and interpretations of the assessments need to be addressed, *including both (a) negative and positive consequences and (b) intended and plausible unintended consequences. (p.2)*
I.2 What is the sampling frame (if any) from which the partipants are chosen?	Explicitly stated (please specify) *elementary and middle schools covered by the MSPAP*
I.3 Which method does the study use to select people, or groups of people (from the sampling frame)?	Explicitly stated (please specify) *A stratified random sampling procedure was used to select schools for the study.* *As a way to implement this procedure, the schools in the population were* *first classified into one of three categories on the basis of their percentile rankings* *for the free or reduced-price lunches participation variable and the* *MSPAP performance gains variable: lower third, middle third, and upper third.* *On the basis of a cross-classification of these two variables (nine combinations* *or cells), the percentage of schools reflecting combinations of the variables varied* *from approximately 8% to 14%. Elementary and middle schools reflecting* *each of the combinations were then randomly sampled. More elementary* *schools were selected because they have fewer teachers per grade than those in* *the middle schools.*
I.4 Planned sample size	Explicitly stated (please specify) *The final samples for this study consisted of 86 of 90 schools for mathematics and language arts (reading and writing), 116 of 161 schools for science,* *and 111 of 161 schools for social studies.* *More elementary* *schools were selected because they have fewer teachers per grade than those in* *the middle schools.Additional schools were randomly selected as potential replacements for schools whose administrators declined the offer to participate. Finally, because contact with schools regarding their participation* *in the study could not be initiated until January 1997, the sample size for the* *1996-1997 instructional year was smaller than the sample size for the* *1998-1999 instructional year.*

	School sample for mathematic
I.5 How representative was the achieved sample (as recruited at the start of the study) in relation to the aims of the sampling frame?	High (please specify) *The achieved sample was representative as recruited at the start of the study in relation to the aims of the sampling frame.*
I.6 If the study involves studying samples prospectively over time, what proportion of the sample dropped out over the course of the study?	Not applicable (not following samples prospectively over time)
I.7 For studies that involve following samples prospectively over time, do the authors provide any information on whether, and/or how, those who dropped out of the study differ from those who remained in the study?	Not applicable (not following samples prospectively over time)
I.8 If the study involves following samples prospectively over time, do authors provide baseline values of key variables, such as those being used as outcomes, and relevant socio-demographic variables?	Not applicable (not following samples prospectively over time)

Section J: Methods - recruitment and consent

J.1 Which methods are used to recruit people into the study?	Implicit (please specify) *The data relevant to this study were obtained from questionnaires that were developed for teachers and students.* *Questionnaires specific to the different subject areas were developed for elementary school (third- and fifth-grade) and middle school (eighth-grade) teachers and students. Teachers completed questionnaires prior to administering the MSPAP, whereas students completed questionnaires within 2 weeks following the administration of the MSPAP.*
J.2 Were any incentives provided to recruit people into the study?	Not stated/unclear (please specify)
J.3 Was consent sought?	Not stated/unclear (please specify)

Section K: Methods - Data Collection

K.1 Which variables or concepts, if any, does the study aim to measure or examine?	Explicitly stated (please specify) *classroom instruction and assessment practices, student learning and motivation, students' and teachers' beliefs about and attitudes toward the assessment, and a school characteristic.*
K.2 Please describe the main types of data collected and specify if they were used to (a) to define the sample; (b) to measure aspects of the sample as findings of the study?	Details *The main types of data are collected through questionnaires to teachers and students. And also through teachers' sample of instruction and assessment tasks that were representative of their classroom materials across the school year. The data collected were used to measure aspects of the sample as findings of the study.*
K.3 Which methods were used to collect the data?	Self-completion questionnaire
K.4 Details of data collection intruments or tool(s).	Explicitly stated (please specify) *The questionnaires consisted of both Likert-type (usually 4-point scales) and constructed response items. Some questions pertaining to the support for the MSPAP and the beliefs about the MSPAP were based on a previous study in which the consequences of state assessments were examined (Koretz, Mitchell, Barron, & Keith, 1996).* *Sets of items on the teacher questionnaire were combined and validated by using confirmatory factor analytic (CFA) methods and measures of internal consistency to reflect the dimensions described in the following list (e.g., Lane, Stone, Parke, Hansen, & Cerrillo, 2000). For the purposes of the CFA, subsets of items were combined to form two indicators for each dimension (about equal numbers of items) except the MSPAP Impact dimension. For this dimension, only one indicator was used because the items could not be divided into two meaningful components. The indicators within each dimension are also given in the following list, as is the total number of items for each dimension, which varied to a small degree across content areas:* *• MSPAP Familiarity (~6 items): (a) teachers' general familiarity with the MSPAP (i.e., purpose, format); (b) teachers' familiarity with MSPAP results (i.e., ability to interpret, use, and explain results)* *• MSPAP Support (~7 items): (a) teachers' general support for the MSPAP*

	(i.e., holding schools accountable, beliefs); (b) teachers' support of the MSPAP for instructional purposes • Current Instruction and Assessment in the Classroom (~20 items): (a) degree to which instruction and assessment reflected each of the state-defined learning outcome standards; (b) extent to which instruction and assessment reflected reform-oriented problem types • Change in Instruction and Assessment from 1993 to 1997–1998 (~20 items): (a) change in emphasis on learning outcomes; (b) change in emphasis on the use of reform-oriented problem types • MSPAP's Impact on Classroom Instruction and Assessment (~6 items): (a) extent to which the MSPAP influenced changes in the classroom • Professional Development Support for MSPAP (~8 items): (a) types of activities related to the MSPAP; (b) amount of professional development support *The student questionnaires paralleled the teacher questionnaires when appropriate. Thus, students were also asked about the nature of instruction and classroom assessment activities.* *So that individual differences in change can be modeled and the correlates of change can be assessed, two levels of statistical modeling are required (e.g., Bryk & Raudenbush, 1992). In the first level, Level 1 (within-school model), trends across the repeated measurements for individual schools are modeled. At Level 2 (between-schools model), individual differences in change across schools from Level 1 are modeled in relation to hypothesized explanatory factors (e.g., dimensions from the teacher and student questionnaires, and the variable of percentage of free or reduced-price lunches).* *Dimensions underlying the questionnaires administered to teachers and students from the schools in the sample were hypothesized to explain individual differences in school performance with time. However, because of the relatively small number of schools in the sample, this study focused on a subset of dimensions from the questionnaires that may be of more interest to stakeholders. From the teacher questionnaire, two dimensions were*

	examined: (a) MSPAP Impact and (b) Current Classroom Instruction and Assessment. The MSPAP Impact dimension, as opposed to the Change in Classroom Instruction and Assessment dimension, was used because it focused more explicitly on the impact of the MSPAP. The Change in Classroom Instruction and Assessment dimension assessed the classroom environment more generally. From the student questionnaire, the Current Instruction dimension and two Likert-type scaled items were analyzed: (a) In class this year, how often did you work on tasks such as those on the MSPAP? (b) How important is it for you to do well on the MSPAP?
K.5 Who collected the data?	Researcher Coding is based on: Reviewers' inference
K.6 Do the authors' describe any ways they addressed the repeatability or reliability of their data collection tools/methods?	*Details* *The instruments (The questionnaires consisted of both Likert-type (usually 4-point scales) and constructed response items) were piloted in the spring of 1996 in Maryland schools and were reviewed by Maryland teachers. Teachers reviewed the questionnaires to ensure that the important outcomes and processes were being measured. There is an element of triangulation by asking the same questions to different types of informants (teachers, students) when this was possible e.g the student questionnaires paralleled the teacher questionnaires when appropriate.*
K.7 Do the authors describe any ways they have addressed the validity or trustworthiness of their data collection tools/methods?	Details *The questionnaires consisted of both Likert-type (usually 4-point scales) and constructed response items. Some questions pertaining to the support for the MSPAP and the beliefs about the MSPAP were based on a previous study in which the consequences of state assessments were examined (Koretz, Mitchell, Barron, & Keith, 1996). The instruments were piloted in the spring of 1996 in Maryland schools and were reviewed by Maryland teachers. Teachers reviewed the questionnaires to ensure that the important outcomes and processes were being measured.*
K.8 Was there a concealment of which group that subjects were assigned to (i.e. the intervention or control) or other key factors from those carrying	No (please specify) *There is not any need to concealment. he following*

out measurement of outcome - if relevant?	*variables were examined through questions in the questionnaire:* *classroom instruction and assessment practices, student learning and motivation, students' and teachers' beliefs about and attitudes toward the assessment, and a school characteristic.*
K.9 Where were the data collected?	Educational Institution (please specify)

Section L: Methods - data analysis

L.1 What rationale do the authors give for the methods of analysis for the study?	Details *Random coefficient* *or growth models were used to examine MSPAP performance from 1993* *to 1998 in relation to variables derived from the teacher and student questionnaires,* *and the school characteristic of percentage of free or reduced-price* *lunches, which served as a proxy for SES. These methodologies are* *particularly well suited for studying processes in which change is considered to* *be continuous but individual differences occur in the pattern of change (i.e.,* *initial level and rate of change). Further, these methodologies allow identification* *of factors that affect the patterns of change. This type of analysis cannot be* *modeled by time-specific comparisons involving group-level (i.e., means)* *differences*
L.2 Which methods were used to analyse the data?	Explicitly stated (please specify) *Sets of items on the teacher questionnaire were combined and validated by using* *confirmatory factor analytic (CFA) methods and measures of internal consistency* *to reflect the dimensions described in the following list (e.g., Lane, Stone,* *Parke, Hansen, & Cerrillo, 2000). For the purposes of the CFA, subsets of items* *were combined to form two indicators for each dimension (about equal numbers* *of items) except the MSPAP Impact dimension. For this dimension, only one indicator* *was used because the items could not be divided into two meaningful components.* *The indicators within each dimension are also given in the following list,* *as is the total number of items for each dimension, which varied to a small degree*

	across content areas: *• MSPAP Familiarity (~6 items): (a) teachers' general familiarity with the* *MSPAP (i.e., purpose, format); (b) teachers' familiarity with MSPAP results* *(i.e., ability to interpret, use, and explain results)* *• MSPAP Support (~7 items): (a) teachers' general support for the MSPAP* *(i.e., holding schools accountable, beliefs); (b) teachers' support of the* *MSPAP for instructional purposes* *• Current Instruction and Assessment in the Classroom (~20 items): (a) degree* *to which instruction and assessment reflected each of the state-defined learning* *outcome standards; (b) extent to which instruction and assessment reflected* *reform-oriented problem types • Change in Instruction and Assessment from 1993 to 1997–1998 (~20 items):* *(a) change in emphasis on learning outcomes; (b) change in emphasis on the* *use of reform-oriented problem types* *• MSPAP's Impact on Classroom Instruction and Assessment (~6 items): (a)* *extent to which the MSPAP influenced changes in the classroom* *• Professional Development Support for MSPAP (~8 items): (a) types of activities* *related to the MSPAP; (b) amount of professional development support* *Across the content areas, CFA results supported the different dimensions just* *specified. Nonsignificant chi-square statistics for models reflecting the preceding* *structure were found.*
L.3 Which statistical methods, if any, were used in the analysis?	Details *In this study, the growth models were estimated by using the SEM program AMOS*
L.4 Did the study address multiplicity by reporting ancillary analyses, including sub-group analyses and adjusted analyses, and do the authors report on whether these were pre-specified or exploratory?	No (please specify) *the study does not address multiplicity by reporting ancillary analyses, including sub-group analyses and adjusted analyses.*
L.5 Do the authors describe strategies used in the analysis to control for bias from confounding variables?	No
L.6 For evaluation studies that use prospective allocation, please specify the basis on which data analysis was carried out.	Not applicable (not an evaluation study with prospective allocation)

L.7 Do the authors describe any ways they have addressed the repeatability or reliability of data analysis?	*Details* *Nonsignificant chi-square statistics for models reflecting the preceding structure were found. In addition, these models significantly improved the model-data fit afforded by lower order models (i.e., 1-factor models). Finally, measures of internal consistency (coefficient ¤) for the separate indicators of the dimensions ranged from .68 to .94 across the content areas.*
L.8 Do the authors describe any ways that they have addressed the validity or trustworthiness of data analysis?	Details *Sets of items on the teacher questionnaire were combined and validated by using confirmatory factor analytic (CFA) methods and measures of internal consistency to reflect the dimensions.*
L.9 If the study uses qualitative methods, how well has diversity of perspective and content been explored?	Details *not a qualitative study*
L.10 If the study uses qualitative methods, how well has the detail, depth and complexity (i.e. the richness) of the data been conveyed?	Details *not a qualitative study*
L.11 If the study uses qualitative methods, has analysis been conducted such that context is preserved?	Details *not a qualitative study*

Section M: Quality of study - reporting

M.1 Is the context of the study adequately described?	Yes (please specify) *Yes. There is no special reason for conducting the study at the time it was done, except data were available and the questions are important. The study is tightly connected to relevant research, data provided by the Maryland State Department of Education, study funded by a grant from US Dept. of Ed., and the study carried out during 1996- (probably 1998 or 1999).*
M.2 Are the aims of the study clearly reported?	Yes (please specify) *Yes. The broad aims are to give more comprehensive and direct evidence for the consequences of statewide assessments programmes than previously research.* *The study will especially explore the relationships between changes in the scores from the MSPAP from 1993 to 1998 and teacher, student, and school variables.* *No specific hypothesis is expressed.*

M.3 Is there an adequate description of the sample used in the study and how the sample was identified and recruited?	Yes (please specify) *Yes. The study has a thorough description of the sample, which is gathered through a stratified random sampling procedure. Profound explanation of the representativeness of the samples.* *Participants consent to participate by answering.*
M.4 Is there an adequate description of the methods used in the study to collect data?	Yes (please specify) *Yes. The scores for schools and proxys for SES were provided by personnel within the Maryland State Department of Education. Other data come from teacher and student questionnaires, but here we get no information on who collected the data or where.*
M.5 Is there an adequate description of the methods of data analysis?	Yes (please specify) *Yes. Schools were placed in a three times three grid based on MSPAP performance gain (lower third, middle third, upper third) and SES variables (lower third, middle third, upper third). The items in the questionnaires were grouped to form indicators for six dimensions using confirmatory factor analytic methods and measures of internal consistency.* *The authors suggest a model including differences within schools.* *Bias due to school size is discussed.*
M.6 Is the study replicable from this report?	No (please specify) *No. The full questionnaires are not available.*
M.7 Do the authors state where the full, original data are stored?	No (please specify)
M.8 Do the authors avoid selective reporting bias? (e.g. do they report on all variables they aimed to study, as specified in their aims/research questions?)	Yes (please specify) *Yes. All items in the questionnaires seem to be used (but it is not completely clear).*

Section N: Quality of the study - Weight of evidence

N.1 Are there ethical concerns about the way the study was done?	No (please specify) *No. We get no information on whether the parents to the students in the elementary schools have given consent, but the items are not dealing with sensitive matters.* *The funding is not likely to have influenced the study.*
N.2 Were users / relatives of users appropriately involved in the design or conduct of the study?	Yes, a lot (please specify) *Yes, a lot. As stated in N1 no consent seems neces-*

	sary. Teachers reviewed the questionnaires to ensure that the important outcomes and processes were being measured.
N.3 Is there sufficient justification for why the study was done the way it was?	Yes (please specify) *Yes. The methods are relevant for the aims of the study, the study is well linked to existing research, and it is performed at a time where changes in test performance are measurable. The danger is that the study links teachers focus on current perceptions with a long term growth rate.*
N.4 Was the choice of research design appropriate for addressing the research question(s) posed?	yes, completely (please specify) *Yes, completely. The study compares changes in test scores during a period with classroom and school variables, which seems an adequate way to monitor consequences of a program over time. It would have gained by taking into account the differences within the individual school.*
N.5 Have sufficient attempts been made to establish the repeatability or reliability of data collection methods or tools?	No, none (please specify) *No. We get no insight in the way the register date or the questionnaires are collected.*
N.6 Have sufficient attempts been made to establish the validity or trustworthiness of data collection tools and methods?	Yes, some attempt (please specify) *Yes. The authorities are supposed to give the right data, and the relevant teachers were reached through the schools.*
N.7 Have sufficient attempts been made to establish the repeatability or reliability of data analysis?	Yes (please specify) *Yes. The study uses standard tools and discusses its use in great detail. The authors discuss possible consequences of sampling errors.*
N.8 Have sufficient attempts been made to establish the validity or trustworthiness of data analysis?	Yes, good (please specify) *Yes, good. Authors discuss problems with sample sizes and with comparing an assessment program that have been in place for several years with current teacher perceptions.*
N.9 To what extent are the research design and methods employed able to rule out any other sources of error/bias which would lead to alternative explanations for the findings of the study?	A little (please specify) *A little. As the authors point out themselves, the used model ignores important differences that may exist among teachers (and classes) within schools. These differences might be bigger than the differences between schools.*
N.10 How generalisable are the study results?	Details *The study results seem generalisable. The findings are not surprising and to a large degree explainable.*

N.11 In light of the above, do the reviewers differ from the authors over the findings or conclusions of the study?	Not applicable (no difference in conclusions) *No difference.*
N.12 Have sufficient attempts been made to justify the conclusions drawn from the findings, so that the conclusions are trustworthy?	High trustworthiness
N.13 Weight of evidence A: Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)?	High trustworthiness
N.14 Weight of evidence B: Appropriateness of research design and analysis for addressing the question, or sub-questions, of this specific systematic review.	Medium
N.15 Weight of evidence C: Relevance of particular focus of the study (including conceptual focus, context, sample and measures) for addressing the question, or sub-questions, of this specific systematic review	Medium *The study focuses on school growth rates in test scores and tries to explain these with school and classroom variables*
N.16 Weight of evidence D: Overall weight of evidence	Medium

7.2 DEC 1 Review specific extra questions

Section A: Test and testing characteristics

A.1 What is the overall purpose of testing?	Measuring students' level of achievement in a subject/domain Grading students or comparing students' level of achievement to that of others Informing instructional practises *MSPAP is intended to promote performance-based instruction.* Accountability/monitoring educational quality
A.2 Which competences/skills are adressed in the test	Procedural skills
A.3 What is the task mode of the test?	Open *In the study it is mentioned "the MSPAP requires students to develop written responses to interdisciplinary tasks" from this the reviewer infer that the questions may be open.*

A.4 How is the response format of the test?	Written (extended) *The reviewer's inference.* Other (please specify) *Unknown*
A.5 Who is scoring the test?	Other (please specify) *nothing is mentioned*
A.6 Who makes use of/are the main audience for results?	Local teachers Head master /school board Local authority
A.7 What are the stakes?	For students (please specify)) *No stakes at all. results only developed at school level* For others (please specify) *unknown*
A.8 If a scale is developed and used for reporting, what does it refer to?	No scale involved or no scale mentioned *It refers to schools. and not to students*
A.9 What is the format for communication the results to students?	Not applicable *Unknown*

Section B: Direction of test effect/impact/influence

B.1 Are sociological issues adressed in relation to the effective use of testing and outcomes?	No
B.2 Is the effective use of the test seen in relation to student characteristics?	No
B.3 Is the effective use of the test seen in relation to teacher characteristics?	No
B.4 Is the effective use of the test seen in relation to the instructional setting?	No

8 Appendiks 3: Abstract af 43 undersøgelser omtalt i kapitel 4

Nedenfor gengives et abstract af de 43 undersøgelser, der indgår i de narrative synteser, gengivet i afsnit 4.2. Ved hvert abstract er det angivet, om undersøgelsen bidrager med høj, hhv. medium evidens til syntesedannelsen. Denne vurdering er baseret på de principper, som er gengivet i afsnit 3.6. Det bør noteres, at de to vurderinger vedrører de enkelte undersøgelsers bidrag til belysning af dette systematiske reviews problemstilling og ikke studiernes forskningskvalitet i al almindelighed. Efter studierne er anført, hvilke relationer i den konceptuelle model, jf. Figur 1.1, side 32, primært belyser.

Alderson, J. C.; Hamp-Lyons, L. (1996) TOEFL preparation Courses: A study of Washback

Studiet har til formål at undersøge washback effekter af TOEFL (Test of English as a Foreign Language). Studiet er en etnografisk undersøgelse, designet som et case-control studie. Undersøgelsen konstaterer, at der kan påvises en række forskelle mellem henholdsvis TOEFL og ikke-TOEFL klasser. Blandt undersøgelsens resultater kan fremhæves, at lærere i TOEFL klasser typisk har mere taletid end elever i ikke-TOEFL klasser, at der er mindre latter i TOEFL klasser, og at der bruges mindre tid på gruppearbejde. Imidlertid konkluderes det, at disse forskelle ikke alene skal ses som en virkning af testen, men i lige så høj grad kan tilskrives administration, undervisningsmateriale, klassestørrelse og lærerne.

Relation: 4, 5

Andrews, S.; Fullilove, J.; Wong, Y. (2002) Targeting Washback – A case study

Studiets formål er at undersøge washback effekterne af en mundtlig eksamen i engelsk som andetsprog, introduceret som et supplement til en eksisterende skriftlig eksamen i Hong Kong. Studiets design er quasi-eksperimentelt. Undersøgelsen viser, at der er små forskelle i testresultaterne mellem de elever, som blev udsat for den mundtlige eksamen, og de elever som alene blev udsat for en skriftlig eksamen. Det konkluderes dog, at forskellene mellem de to elevgrupper ikke kan betragtes som et udtryk for, at den mundtlige eksamensform bidrager til en reel forbedring af elevernes færdigheder, men snarere som et udtryk for at eleverne har tilpasset sig testen.

Relation: 4, 5

Bauer, S. et al. (1990) Controlling Curricular Change through State-mandated testing: Teacher's Views and Perceptions

Studiet undersøger lærernes opfattelse af Program Evaluation Tests (PET), indført i staten New York. Programmet vedrører natur- og samfundsfag i hhv. fjerde og sjette klasse. Undersøgelsen er designet som et views study. Undersøgelsen viser, at selvom PET ikke havde til hensigt at vurdere individuelle elever eller lærere, opfattedes testen alligevel som high-stakes og lærerne erklærede, at de indrettede undervisningen med henblik på forberedelse til testningen.

Relation: 2, 4

Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "On-Line" Test Feedback on the Seriousness of Subsequent Errors

I undersøgelsen fik tre grupper elever en matematisk test. Når et item var besvaret forkert, fik eleverne i den første gruppe meddelt, at svaret var forkert, eleverne i den anden gruppe, hvad det rigtige svar er, mens eleverne i den tredje gruppe fik meddelt, hvilken regneregel der skulle være anvendt og hvori fejlen bestod. Der er anvendt et RCT-design. Undersøgelsen opdeler fejlbesvarelser i seriøse og ikke seriøse fejl. Ved de seriøse fejl kunne det ikke påvises, at feedback meddelelsen gjorde nogen forskel for elevers senere præstation i en efterfølgende test, mens dette var tilfældet ved ikke-seriøse fejl. Konklusionen er, at selv den mest informative af de anvendte feedbackmetoder er ineffektiv som pædagogisk metode, når fejltypen er seriøs.

Relation: 2

Boesen, J. (2006) National Course tests' impact on teachers' written classroom assessment

Formålet med studiet er at undersøge forholdet mellem nationale test (National Core Test, NCT) og læreres selvudviklede test (Teacher Made Test, TMT). Studiet er primært dokumentbaseret. Undersøgelsen viser, at langt størstedelen af lærernes selvudviklede test ikke modsvarer de nationale test. Det konkluderes derfor, at de nationale test ikke fungerer efter hensigten, dvs. som eksemplariske modeller for udvikling af lærernes egne test, og dermed et redskab, der sikrer at lærernes instruktion sker i overensstemmelse med de nationale læseplaner og curriculum.

Relation: 4

Chen, L.M. (2002) Washback of a Public Exam on English Teaching

Studiet har til formål at undersøge washback effekterne af en ny offentlig eksamination i engelsk, som blev introduceret i Taiwan i fællesskab med en fornyelse af curriculum samt nye obligatoriske lærebøger. Undersøgelsen er udført som et views study. Studiet viser, at den nye eksamination førte til en indsnævring af både curriculum og instruktion. Således indrettede lærerne undervisningen med henblik på forberedelse til eksamen i stedet for at ændre deres instruktion i overensstemmelse med politikernes intentioner. Det konkluderes, at intentionen om at tilvejebringe positive washback effekter via ændringer i henholdsvis eksamensform og indhold er vanskelig at realisere.

Relation: 4

Cheng, L. W., Y.; Curtis, A. (2004) The Washback Effect of a Public Examination Change on Teachers' Perceptions toward Their Classroom Teaching

Studiet undersøger hvorvidt omfattende ændringer i et eksisterende high-stake testsystem i Hong kong, resulterer i tilsvarende ændringer i læreres opfattelse af undervisning og læring i faget Engelsk. Studiet har gennem en spørgeskemaundersøgelse spurgt til lærernes opfattelser af deres undervisning de første år efter indførelse af det nye testsystem. Studiet finder frem til, at selvom lærerne efterhånden synes at have nogenlunde overenstemmende opfattelser med politikkerne, er lærernes daglige aktiviteter og deres grundlæggende opfattelser af læringstilgange, knyttet an til det nye testsystem, stadig uændrede. Forfatterne konkluderer, at ændringer i testsystemer ikke alene er tilstrækkelige til at realisere intenderede mål med undervisningen.

Cheng, L., Rogers, T., & Hu, H. (2004) ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures

Målet med undersøgelsen er at sammenligne læreres brug af test på tværs af tre lande, henholdsvis Canada, Hong Kong og Beijing. Studiet finder frem til, at lærere vælger at anvende test/testresultater af forskellige grunde, som viser sig at have en del med undervisningskulturen at gøre. Især adskiller lærerne i Beijing sig fra lærerne i Canada og Hong Kong. Anvendelse af standardiserede tests er i højere grad gældende for lærerne i Beijing (8 ud af 10), end for Canada (3 ud af 10) og Hong Kong (1 ud af 10). Især i Beijing hentes test på nettet. Lærerne i de tre lande er dog enige om, at brugen af testdata har elev-centrerede formål, såsom at opnå information om de studerendes fremskridt, at kunne give feedback til de studerende om deres fremskridt, at diagnosticere styrker og svagheder hos de studerende, at fastsætte karakterer for de studerende.

Danmarks Evalueringsinstitut (2002). Folkeskolens afgangsprøver. Prøvernes betydning og sammenhæng med undervisningen

Studiet er primært en spørgeskema undersøgelse bland**. Formålet med studiet er en evaluering af folkeskolens afgangsprøve. Studiet undersøger både virkninger af afgangsprøverne på elever samt lærer og sammenhængen mellem undervisningens tilrettelæggelse og prøverne. Studiet finder bl.a. frem til, at test har en motiverende effekt på lærer og elev og at test er med til at styrke form og indhold i faget. Eleverne er dog mindre involverede og klasseundervisning er det mest almindelige. Danmarks Evalueringsinstitut anbefaler, at test introduceres i flere fag i folkeskolen.

Doran, H. C. (2001) Evaluating the Consequential Aspect of Validity on the Arizona Instrument to Measure Standards

Formålet med studiet er at undersøge, hvorvidt de positive effekter forbundet med anvendelse af high stakes test (Arizona Instruments to Measure Standards, AIMS) i staten Arizona var gældende for alle klassetrin – og ikke alene de klassetrin, som blev udsat for test. Undersøgelsen er baseret på spørgeskema data fra lærere. Resultaterne fra studiet viser, at de positive effekter forbundet med testning var signifikant større i de klasser, som blev udsat for test end i de ikke-testede klasser. Det konkluderes på denne baggrund, at virkningerne af testen ikke lever op til de politiske intentioner.

Relation: 5 (NB! Måske forkert brug i brødteksten??!!)**

Ferman, I., Watanabe, Y. & Curtis, A. (2004) The Washback of an EFL National Oral Matriculation Test to Teaching and Learning

Studiet har til formål at undersøge washback effekterne af en mundtlig immatrikulationstest i engelsk som fremmedsprog. Studiet er gennemført som en tværsnitsundersøgelse, hvori der indgår data fra tre forskellige skoletyper i Israel. Det konkluderes, at den mundtlige test resulterer i stærke - og som oftest negative - washback effekter på undervisning og læring. Lærerne indretter en stor del af undervisningstiden med henblik på forberedelse til testningen, elevernes indlæring præges i højere grad af udenadslære, og såvel elever som lærere forbinder testningen med angst. Undersøgelsen viser dog, at skoleinspektørerne samt et mindretal af lærerne er positivt stemte over for testen og dens (potentielle) evne til at generere de tilsigtede washback effekter.

Relation: 4, 5

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004) Test Preparation in New Jersey: Inquiry-Oriented and Didactic Responses

Studiet undersøger mulige virkninger på undervisning tre år efter indførelsen af en standardiseret test i matematik og videnskab - Elementary School Performance Assessment (ESPA) - for niårige i New Jersey. Forfatterne hævder, at ingen af deres forskningsresultater er signifikante, men peger generelt på, at indretning af undervisningen med henblik på forberedelse til test er mere udpræget i skoledistrikter med overvejdende lavt præsterende elever. Pres på skolen om bedre testresultater, resulterer i yderligere intens forberedelse til test. Og jo mindre viden lærerne har om testen, des mere begrænsede bliver deres undervisningsmetoder.

Fuchs, L. S., & et al. (1989) Monitoring Reading Growth Using Student Recalls: Effects of Two Teacher Feedback Systems

Fuchs & al. sammenligner to feedbacksystemer inden for læsetest. Det ene system registrerede, hvor mange ord eleverne kunne gengive, mens det andet desuden gav læreren information i form af en kvalitativ analyse af de historiekomponenter, der indgik i elevernes gengivelser. Et case-control design er taget i anvendelse. Det viser sig, at feedbacktypen med kvalitativ beskrivelse af elevernes testresultater giver nyttig viden til lærernes udvikling af egen undervisning. Undersøgelsen viser videre, at jo mere information lærere har, des mere nyttig viden får de.

Relation: 1, 2

Goldberg, G. L., & Roswell, B. S. (1999) From perception to practice: The impact of teachers' scoring experience on performacne-based instruction and classroom assessment. Educational Assessment

Goldberg & Roswell undersøger læreres instruktion og anvendelse af den specifikke test, the Maryland School Performance Assessment Program (MSPAP), der indebærer test i form af praktiske opgaver. Undersøgelsen er et etnografisk studie med spørgeskemabesvarelser, interviews og observationer med udvalgte lærere, samt analyser af artefakter. Studiet konkluderer, at lærere finder testen vigtig i og med, at de mener, testningen gør dem til endnu mere reflekterende, kritiske og velovervejede lærere. Dog viser analyser af datamaterialet, at lærere kun havde tilegnet sig færdigheden i at undervise i de praktiske opgaver på baggrund af testen på en overfladisk og ufuldstændig måde.

Relation: 4

Gutkin, J. (1985). The Effect of Diagnostic Inservice Training on Class Reading Achievement and the Number of Lessons Covered

Studiet tager udgangspunkt i en serie kriterium-relaterede test med efterfølgende undervisning i læsning, talt sprog og aritmetik. På baggrund af disse resultater får en gruppe lærere løbende assistance af en superviser. Superviserne assisterer lærerne i, hvorledes elevernes testdata anvendes bedst pædagogisk. I kontrolgruppen modtager lærerne ikke assistance. Undersøgelsen kunne ikke påvise, at forsøgsgruppen opnåede bedre testdata end kontrolgruppen. Forfatteren mener, at dette kan ses som en type 2 fejl. Studiet påviser dog, at jo flere fagområder og jo flere timer undervisningen omfatter, des bedre testdata får eleverne på alle undersøgte fagområder. Undersøgelsen kan ikke påvise, at øget information om testdata medfører viden, der kan omsættes i bedre elevlæring/øgede testdata.

Harlen, W., & Ruth Deakin, C. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning

Et systematisk review, der har til hensigt at finde evidens for om summative test har en indvirkning på elevers motivation for at lære. Studiet har en bred tilgang, der også indbefatter virknin-

ger på lærernes pædagogiske intervention. 19 studier inkluderes i reviewet. Studiet konkluderer bl.a., at der er stærk evidens for, at lærere anlægger en ensidig undervisningsform (transmission teaching of knowledge), når elever skal bestå en high stakes test. Denne læringsstil favoriserer nogle elever, mens andre elever, der foretrækker mere aktivitet og kreativitet i læringen, oplever lavere selvværd (self-esteem). Efter introduktionen af the National Curriculum Tests i England, havde lavt præsterende elever lavere selvværd end højtpræsterende elever. Før introduktionen af disse tests var der ingen korrelation mellem selvværd og præstationer (achievement). Derudover at der stærk evidens for, at elever oplever lærernes formative test som summative test, uafhængig af lærerens intention, sandsynligvis et resultat af lærerens bekymring (overconcern) for præstationer frem for processer.

Harlen, W., & Crick, R. D. (2003). A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills

Gennem et systematisk review har undersøgelsen til formål at finde evidensen af effekten på lærere og elever, når IT anvendes til at teste elevernes færdigheder i kreativ og kritisk tænkning. Undersøgelsen har herunder desuden til formål at klarlægge positive og negative aspekter ved brugen af dette medie mhp. at bestemme hvilke summative og/eller formative virkninger, der er herved. Harlen & Deakin har om IT som et medium fundet, at studier med høj evidensvægt viser, at IT kan hjælpe lærere ved at lagre og registrere information om, hvordan elever udvikler forståelse af nyt stof og overtage noget af arbejdet med at rette opgaver og give feedback, så læreren kan fokusere på den form for indlæring, som IT-programmerne ikke yder. Studier med medium evidensvægt viser, at computeren både kan teste og give feedback, hvorved elevernes præstationer bliver bedre ved efterfølgende sammenligning med test, der er papirbaserede. Computere er desuden i stand til at give feedback om den proces, hvorigennem eleven når til en løsning. Det øger mængden af feedback til elev og lærer.

Relation: 2, 3a

Henderson, S., Petrosino, A., Guckenburg, S. & Hamilton, S. (2007) Measuring How Benchmark Assessments Affect Student Achievement

Dette studie har til formål at undersøge, hvorledes brug af kvartalsvise benchmark eksaminer indvirker på elevpræstationer. Studiet er gennemført som et quasi-eksperimentelt kohorte studie. Analyserne viser, at der ikke kan påvises nogen signifikante forskelle i elevpræstationer mellem de skoler, der benytter sig af de kvartalsvise benchmark eksaminer, og skoler der ikke benytter denne prøveform.

Relation: 5 (kan benchmark eksaminer oversættes til standpunktseksamen??), dette studie er ikke kommenteret selvstændigt på i brødteksten SE: måske kvartalprøver er bedre.**

Higgins, N., & Rice, E. (1991). Teachers' Perspectives on Competency-Based Testing

Et etnografisk studie fra USA som undersøger, dels hvilke teknikker lærere bruger i evalueringen af eleverne, dels lærernes holdninger og brug af forskellige test-metoder. Studiet finder frem til, at lærerne foretrækker test, som de kan integrere i deres undervisning. Kommunale kompetence-test blev opfattet som formale, ikke relaterede til lærernes undervisningsplan og ikke anvendelige i evalueringen af de studerendes kompetencer. Forfatterne konkluderer, at i udviklingen af testprogrammer må læreren i højere grad medtænkes, således at test mere hensigtsmæssigt kan integreres i lærernes planlægning og undervisning.

Howe, M. E., & Thames, D. G. (1996) Mississippi Reading Teachers' Perceptions toward the Interpretation of Results from Reformed Standardized Assessment in Mississippi

Dette studie undersøger læseunderviseres opfattelse af tolkningen af resultater af standardiserede test i Mississippi (Mississippi Assessment System, MAS). Studiet er udført som et views stu-

dy. Lærernes svar siger, at effektiv anvendelse af data indsamlet fra statens standardiserede test kan hjælpe lærerne til at vælge passende undervisningsstrategier og muligvis også at designe curriculum og uddannelsesprogrammer.

Relation: 2 (NB! referencen skal indplaceres i brødteksten)**

Jia, Y., Eslami, Z. R., & Burlbaw, L. M. (2006). ESL Teachers' Perceptions and Factors Influencing Their Use of Classroom-Based Reading Assessment

Et etnografisk studie fra USA, der undersøger underviseres opfattelse og antagelser om forskellige typer af test. Studiet finder frem til, at lærerne generelt bruger tre typer af test, lærernes egne test, observation i forbindelse med undervisningen, formelle test, nationale test og computerbaserede læsetest. Lærerne udtrykker en generel bekymring omkring enhver form for formel skriftlig test. Mange elever er ikke vant til test og føler sig nervøse og skuffede under testen eller de forsøger at ignorere test og gør ikke deres bedste. Derfor føler lærerne, at det er svært at få en korrekt evaluering af en elevs læsefærdighed. Lærerne kritiserer desuden testen for at ikke at hænge sammen med curriculum. Forfatterne konkluderer, at lærernes egne udviklede test, bør accepteres i langt højere grad af politikere og skoleadministratorer.

Kozulin, A., & Garb, E. (2004). Dynamic Assessment of Literacy: English as a Third Language

Studiets design er en pre-post test udført i Israel men henbilk på at vurdere effektiviteten af Vygotskys sociokulturelle teori og begrebet dynamisk testning. Undersøgelsen finder bl.a., at dynamisk testning er effektiv med hensyn til at give eleverne mere indsigt i effektive kognitive strategier, idet elevernes indsigt heri bliver efterfulgt af øgede testresultater. Studiet finder dog, at de mange forbedringer ikke er ens for erfarne studerende og nyankomne immigranter. Forfatterne konkluderer, at en dynamisk test kan være værdifuld for visse studerende.

Lane, S., Parke, C. S., & Stone, C. A. (2002) The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance

Dette studie har til formål at undersøge, hvorledes the Maryland School Performance Assessment Program (MSPAP) og the Maryland Learning Outcomes (MLOs) indvirker på henholdsvis klasserumsinstruktion, testning, professionel udvikling og elevindlæring i faget matematik. Studiet er designet som et views study. Det konkluderes, at flertallet af lærere betragter MSPAP som et nyttigt redskab til at forbedre undervisningen, og at testen har en positiv effekt på undervisningen. Endvidere v Dynamic Assessment of Literacy: English as a Third Language iser undersøgelsen, at skoler, hvis lærere giver udtryk for, at testprogrammet har stor indflydelse på deres undervisning, kan fremvise en større fremgang i elevernes test-scores over tid.

Relation: 4, 5

Luxia, Q. (2007) Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China

Studiet har til formål at undersøge washback effekterne af en national test indført i Kina i faget engelsk (National Matriculation English Test, NMET). Studiet er gennemført som en tværsnitsundersøgelse. Et af testens primære pædagogiske formål var at styrke elevernes kommunikative færdigheder. Studiet fokuserer på den del af testen, som omhandler skrivning, og konkluderer, at selvom skrivning blev praktiseret i skolerne, skete dette ikke på en måde, der var i overensstemmelse med testens intentioner. Såvel lærere som elever forsømte den kommunikative del af skrivningen og fokuserede i stedet på henholdsvis testningens og karaktergiverens forventede præferencer. Forfatterne slutter på denne baggrund, at high-stakes test ikke er et effektivt redskab til realisering af den type forandring, som politikerne ønskede at fremkalde.

Relation: 4

Mason, R. (2005). The effect of formal assessment on secondary school art and design education.

Denne undersøgelse er et systematisk review, der har til formål at undersøge effekterne af summative tests på læseplanen for billedkunst, lærere og elever på gymnasiet. Et af resultaterne ved dette review er, at kvaliteten af forskningen i billedkunst, samt forskningsafrapporteringen er af ringe kvalitet, der gør det vanskeligt at drage konklusioner på det foreliggende materiale. Mason et al. konkluderer, at der er brug for flere undersøgelser, herunder systematiske undersøgelser, der måler testningens effekter på faget billedkunst.

Relation: Studiet har ikke resultater, der kan bidrag til nogen af de fem relationer

Mattsson, H. (1989). Proven i skolan sedda genom lärarnas ögon.

Mattson beskriver, hvordan skriftlige test bruges i slutningen af grundskolen og i gymnasiet på tværs af fag. Undersøgelsen er en tværsnitsundersøgelse, hvis dataindsamlingen bygger på lærerinterviews og læreres spørgeskemabesvarelser. Lærerne mener, at testresultater hjælper dem med at se, hvad de studerende kan, og til at give karakterer. Lærerne bruger også andre oplysninger om eleverne, såsom en vurdering af deres talent, flid, samarbejdsvilje, interesse. Jo bedre en lærer kender en elev fra undervisninges, des mindre rolle spiller test for karaktergivningen.

Relation: 2.

(NFS) – mangler, NerimanSE: jeg har set et resume et eller andet sted!**

Parke, C. S., Lane, S., & Stone, C.A. (2006) Impact of a State Performance Assessment Program in Reading and Writing

Studiet undersøger, hvorledes et bestemt testprogram (Maryland School Performance Assessment, MSPAP) indvirker på henholdsvis opfattelser, undervisning og elevernes læring inden for læsning og skrivning. Studiet er gennemført som en tværsnitsundersøgelse baseret på data indsamlet via spørgeskemaer. Et vigtigt resultat af undersøgelsen er, at testen ifølge underviserne har en tydelig indflydelse på undervisningen. Endvidere konkluderer forfatterne at skoler, der tilpasser deres undervisning til statens reformorienterede mål har opnået bedre resultater i læsning og skrivning. Det bemærkes hertil, at den konkrete test udmærker sig ved en række egenskaber, herunder at testen ikke er en hylde-vare test skabt af en ekstern instans, at skolerne kendte testen og støttede dens mål, at testen er teknisk velfungerende og har en høj reliabilitet og validitet, og endelig at testen alene var high-stakes på skoleniveau og ikke på individniveau, hvorfor lærere og elever ikke var ængstelige for testningen.

Relation: 1, 2, 4, 5

Parker, D. L., & Picard, A. J. (1997). Portraits of Susie: Matching Curriculum, Instruction, and Assessment. Teaching Children Mathematics.

Dette studie er en case study af en tosproget elevs præstation ved en formel test i matematik: Curriculum and Evaluation Standards for School Mathematics (NCTM 1989). Studiet adresserer mulige konsekvenser ved brugen af en standardiseret test til måling af elevens matematikfærdigheder. Det første portræt af eleven viser en yderst motiveret elev, der har interesse for at lære matematik, og som kan tænke abstrakt. Det andet portræt, der bygger på elevens præstation ved en standardiseret test, viser, at eleven scorer lavt i testen, muligvis på grund af begrænsede sprogfærdigheder. Studiet konkluderer, at det er læreres ansvar at udvikle en me-

ningsfuld program i matematik ved at matche læseplan, undervisning og testning af elevfærdigheder, hvis elevens fulde potentiale skal udfoldes.

Relation: 1

Ryan, J., & Williams, J. (2000). National testing and the improvement of classroom teaching: can they coexist?

Undersøgelsen har en tese om, at de fejltyper, som børn på 7 hhv. 14 år opviser, når de testes i matematik, kan analyseres af fagdidaktikere ud fra en diagnostisk tilgang med henblik på mere effektiv tilrettelæggelse af undervisningen. Indsigt i de typer fejl, som en elev begår, kan hjælpe læreren til at danne sig en mental model af elevens forståelse af matematik. Det antages, at denne model kan være et redskab til lærerens arbejde med den enkelte elev. Undersøgelsen bygger på analyser af elevernes besvarelser ved en nationale test i matematik. Det blev konstateret, at diagnosen af 7-åriges fejltyper ikke var til stor hjælp, mens diagnosen af de 14-åriges fejltyper kunne tilbyde lærere indsigt i testtyper, som må antages at være til hjælp den enkelte elevs problemer med matematik.

Relation: 1, 2.

Ryan, P. (1994). Teacher Perspectives of the Impact and Validity of the Mt. Diablo Third Grade-Curriculum-Based Alternative Assessment of Mathematics

Gennem spørgeskema og interview vil studiet finde frem til, hvorledes en test: Curriculum-Based Alternative Assessment of Mathematics (CBAAM), påvirker lærernes undervisning og hvorledes lærerne anvender testen i undervisningen. Studiet konstaterer bl.a. at nogle lærere har ændret deres undervisningspraksis med introduktion af testen, mens andre har gjort det allerede før testen introduceres. Lærere, der anvender testdata som baggrund for deres pædagogiske tiltag, gør dette på klasseniveau, ikke på individniveau.

Scott, C. (2007). Stakeholder perceptions of test impact

Et casestudie fra UK, der omhandler læreres syn på anvendeligheden og virkninger af test. Lærerne fortæller i undersøgelsen, at testdata blev brugt til at forudse niveau for år 6 og at monitorere såvel den enkelte elev som hele elevgruppen og skabe baggrund for planlægning både generelt og i forhold til særlige støtteforanstaltninger. Diagnostiske formål viser sig kun at kunne bruges på et generelt, ikke på et individuelt niveau. Testen havde ikke samme værdi på alle skoler, men forskeren fandt, at den havde indflydelse på elevernes selvagtelse**

Shannon, A. J. (1980). Effects of Methods of Standardized Reading Achievement Test Administration on Attitude toward Reading

Studiet undersøger, om en standardiseret test har en skadelig indvirkning på elevers holdning (attitude) til faget (læsning). Studiet har et kvasi-eksperimentalt design. Analysen af dataen viser, at de studerendes holdning til faget er signifikant påvirket af både testmetoden og de efterfølgende testresultater. Studiet konkluderer, at testresultatet giver en positiv holdning til læsning blandt gode læsere, men giver en negativ holdning hos svage læsere.

Relation: 5

Shohamy, E., & et al. (1996). Test Impact Revisited: Washback Effect Over Time

Dette studie har til hensigt at undersøge to nationalt introducerede sprogtests virkninger over tid på fagene arabisk som andetsprog (AsA) og engelsk som fremmedsprog (EsF). Undersøgelsen er udarbejdet i Israel. Dataindsamlingen er baseret på interviews og spørgeskemaer til skoleinspektører, lærere og elever. Undersøgelsen konkluderer, at testens effekt på EsF er øget, men uændret hvad angår AsA. Tilstedeværelsen af testen er synlig i faget EsF for alle. Lærerne i

faget EsF øger mængden af aktiviteter i mundtlighed i undervisningen med henblik på at træne eleverne til at tage testen. Lærerne oplever testresultaterne som et mål på kvaliteten af deres undervisning og føler sig dermed under et testpres.

Relation: 4

Silis, G. (2005b, 2005). Using test information to improve program delivery

Silis undersøger brugen af testdata til at forbedre undervisning og elevlæring med udgangspunkt i et projekt fra 1996 om brug af testdata til at forbedre undervisningen i modersmål og matematik. Studiet er en cohortundersøgelse, der varer over flere år**[hvor mange år?]. Silis fremhæver, at test især er praktisk for lærerne til planlægning, når de endnu ikke kender eleverne godt. Testdata kan anvendes til at få identificeret, hvilke dele af stoffet der ikke blev dækket ind og derved til at belyse problemområder og bidrage til implementering af tiltag, der kan støtte eleverne. Desuden hjælper testdata læreren med at identificere, om der er dele af læseplanen i matematik, som klassen som helhed, grupper af elever eller enkelte elever i klassen har brug for mere undervisning i.

Relation: 2

Smart, M. (2004). An investigation into the consequential validity of the Secondary Entrance Assessment Examination. Florida State University, Tallahassee.

Smart undersøger konsekvenser og effekter af en ny high-stakes national test; SEA (Secondary Entrance Assessment), for 5. klasser i Trinidad og Tobago. SEA-eksaminationen erstatter den tidligere eksamination Common Entrance Examination (CEE), som viste sig at have visse negative wash-back effekter. På baggrund af udsagn fra ca. 100 lærer konstaterer studiet, at testtypen Secondary Entrance Assessment (SEA)i modsætning til det tidligere testsystem (CEE) har haft en række både positive og negative effekter, der fordeler sig forskelligt på de to undersøgte fag, matematik og sprog. Engelsklærerne på secondary school rapporterer, at de studerende i større grad er blevet mere selvhjulpne og kreative. Men SEA- eksamen stiller ifølge sprogfagslærerne, et større krav om at levere differentieret undervisning, og derved kræver flere lærerressourcer i form af effektiv planlægning af undervisning og organisatoriske færdigheder. Yderligere konkluderer studiet, at lærere allokerer mere tid til fag, der skal testes i, på bekostning af fag, der ikke testes i. Testen har forskellige effekter i fagene engelsk og matematik.

Relation: 4, 3b

Smith, P. S., & et al. (1992). The Impact of End-of-Course Testing on Curriculum and Instruction. Science Education

Undersøgelsen har til formål at undersøge effekterne af North Carolina's testprogram på læseplanen og undervisning i faget kemi, herunder en undersøgelse af kemilærernes holdning til testprogrammet. Studiet har et cross sectional design. I spørgeskemabesvarelserne rapporterer over 64 % af lærerne, at de underviser deres elever med henblik på at gøre dem klar til testen, men størstedelen af disse lærere afviser dog at undervise eleverne i specifikt at blive bedre til at tage testen, blot 37,8 % af lærerne bekræfter dette ved interviewene. Lærere skelner derved mellem det, at undervisningen har elementer af at forberede eleverne på testen overfor, at undervisningen alene tilrettelægges mhp. at forberede eleverne til at tage testen. Studiets konklusion er ydermere, at testen har en effekt på læseplansudformningen, samt at lærerne føler sig underlagt et pres i kraft af testens tilstedeværelse.

Relation: 4 (NB! referencen skal indplaceres i brødteksten)**

Stecher, B., Chun, T., & Barron, S. (2004). The Effects of Assessment-Driven Reform on the Teaching of Writing in Washington State

Studiet undersøger virkninger af en test drevet uddannelsesreform i Washington på skoler og klasser, samt hvilke elementer ved reformen, der ved hjælp af et view study blev opfattes som havende størst betydning. Lærere giver udtryk for, at de fleste elementer ved reformen i store træk har en positiv effekt på undervisning og læring. Studiet konkluderer, at der sker en reallokering af undervisningstid. Lærere meddeler, at de har foretaget ændringer i bl.a. tildelingen af tid til skriveøvelser, i vægtningen af specifikke aspekter ved skrivning, deres undervisnings-metoder, og deres elevers læringsaktiviteter. Der bliver undervist mere i fag, eleverne skal testes i, end i fag, eleverne ikke skal testes i. Studiet konstaterer, at de studerende opnår højere testresultater på skoler hvor lærere rapporterer, at de oplever overensstemmelse mellem læreplaner og test. Ligeledes konstaterer studiet, at elevernes testresultater i læsning og matematik er steget signifikant, når lærere rapporterer, at de har forstået testen.

Relation: 4

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables

Studiet har blandt andet til hensigt at undersøge, hvorledes the Maryland School Performance Assessment Program (MSPAP) har medført ændringer i undervisning, evalueringmetoder og elevernes læring, motivation, elev-og lærertænkning og holdning til evaluering. Det er en cohortundersøgelse, baseret på MSPAP test-data fra 1993 til 1998 og spørgeskemadata fra lærere og elever. En af konklusionerne i undersøgelsen er, at skoler med høj andel af elever med højere socio-økonomisk baggrund præsterer signifikant bedre i løbet af perioden end skoler med høj andel af elever med lavere socio-økonomisk baggrund, der ligeledes udviser stor variation i testscore. Skoler, hvor lærerne i større udstrækning anvender opgaver, der er identiske med opgaverne i MSPAP, øger elevernes testpræstationer markant i løbet af 1993 og 1998.

Relation: 4, 5

Sturman, L. (2003). Teaching to the test: science or intuition? Educational Research

Denne undersøgelse har til formål at give en detaljeret beskrivelse af læreres forskellige måder at forberede eleverne på test i naturfag. Studiet har et cross sectional design. Konklusionerne er blandt andre, at 61 lærere ud af i alt 64 lærere (95 %) forbereder deres elever på testen. 32 af de 64 adspurgte lærere mener ligeledes, at testforberedelsen har erstattet nogle af deres andre almindelige aktiviteter i naturfag. Tid allokeret til testforberedelse varierer mellem få dage op til 8 måneder, der blandt andet anvendes til at undervise eleverne i at afkode en test, samt at lære dem specifikke fremgangsmåder til at klare testen med. Små skoler begynder testforberedelserne senere end store skoler, samt bruger færre ressourcer til det end store skoler. Lavtpræsterende klasser starter tidligere på testforberedelse end middel- og højtpræsterende klasser og bruger dermed flere ressourcer på forberedelserne.

Relation: 4

Tresch, S. (2007). Potential Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen. Bern: H.E.P. Verlag

Afhandlingens mål er at undersøge læreres konkrete anvendelse af testdata i deres undervisning i modersmål og matematik. Undersøgelsen er baseret på et one-group-post-test og view study design. Tresch viser, at lærerne generelt anvender testdata i deres undervisning, men det kræver, at de har god forståelse af disse resultater. Kun få lærere havde ikke planlagt nogen form for opfølgnings-tiltag, de øvrige havde planlagt mellem to og fire tiltag. Lærere lader til at være kompetente til at uddrage konkrete mål og tiltag på basis af feedback fra test. De mere erfarne lærere havde planlagt færre tiltag, men gengæld gennemført dem alle. Test virker godt til at sammenligne klasser og især visse test er gode til at skabe refleksion hos lærerne om deres egne

undervisning. Forfatteren anbefaler, at der gives en præcis beskrivelse af testen så læreren ved, hvad der skal måles, at der så vidt muligt laves flere sammenligninger, gives flere eksempler og tabeller med resultater.

Relation: 1, 2

Wall, D., & Alderson, J. C. (1992) Examining Washback: The Sri Lankan Impact Study (NB!! Forkert reference? - mangler 2005 studiet på referencelisten!!??)**

Formålet med studiet er at undersøge virkningerne af en reform af en eksamen i engelsk som andetsprog, introduceret i Sri Lanka. Studiet er gennemført som en etnografisk, dokumentbaseret, kohorte undersøgelse.** [SE??] Den nye eksamensform skulle motivere lærerne til at ændre deres engelskundervisning i overensstemmelse med indholdet i en nyindført lærebog. Undersøgelsen viser, at selvom lærerne rapporterede om ændringer af form og indhold i undervisningen, kunne sådanne ændringer ikke iagttages, når undervisningen blev observeret. Der var med andre ord forskel mellem det lærerne hævdede om deres undervisning, og den måde de faktisk gennemførte den på.

Relation: 4

Watanabe, Y. (2004) Teacher Factors Mediating Washback

Målet med undersøgelsen er at identificere en række lærerfaktorer, der er medvirkende til testens washback. Gennem interviews og observations finder forfatteren frem til, at undervisningen for nogle læreres vedkommende er blevet ramt af testens (negative) washback – mens det ikke er tilfældet for andre lærere. Lærer-psykologiske faktorer lader dog til at have indflydelse på graden af washback samt skolens kultur.

**

Winfield, L. F. (1987). The relationship between minimum competency testing programs and students' reading proficiency. Implications from the 1983-84 National Assessment of Educational progress in reading and writing

**

Winter, J., Fitz, J., & Firestone William, A. (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales

**

9 Samlet oversigt over kortlagte undersøgelser

En reference, der er markeret med *, har ikke foreligget i tide til, at den har kunnet indgå i genbeskrivelsen.

Afflerbach, P., & Moni, K. (1996). Improving the Usefulness and Effectiveness of Reading Assessment. Instructional Resource No. 33 (No. ED400516).

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL Preparation Courses: A Study of Washback. Language Testing, 13(3), 280-297.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting Washback--A Case Study. System, 30(2), 207-223.

Bauer, S., & et al. (1990). Controlling Curricular Change through State-Mandated Testing: Teacher's Views and Perceptions (No. ED340740).

Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "On-Line" Test Feedback on the Seriousness of Subsequent Errors. Journal of Educational Measurement, 24(2), 145-155.

Boesen, J. (2006). National course tests' impact on teachers' written classroom assessment. Umeå: Department of Mathematics and Mathemical Statistics, Umeå Universitet.

Burnett, F. (1987). Diagnostic and Prescriptive Preparation for the Florida Student State Assessment Test (No. ED317564).

Burnham, B. (1983). Use of Standardized Achievement Test Results in YRBE Elementary Schools (No. ED242791).

Chen, L.-M. (2002). Washback of A Public Exam on English Teaching (No. ED472167).

Cheng, L. (1995). How Does Washback Influence Teaching? Implications for Hong Kong (No. ED385143).

Cheng, L. (1997). How Does Washback Influence Teaching? Implications for Hong Kong. Language and Education, 11(1), 38-54.

Cheng, L. (1999). Changing Assessment: Washback on Teacher Perceptions and Actions. Teaching and Teacher Education, 15(3), 253-271.

Cheng, L. (2003). Looking at the Impact of a Public Examination Change on Secondary Classroom Teaching: A Hong Kong Case Study. Journal of Classroom Interaction, 38(1), 1-10.

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. Teacher Assessment., 21(3), 360-389.

Cheng, L. W., Y.; Curtis, A. (2004). The Washback Effect of a Public Examination Change on Teachers' Perceptions toward Their Classroom Teaching. MT: Monograph Title Washback in language testing: Research contexts and methods.

Danmarks Evalueringsinstitut (2002). Folkeskolens afgangsprøver. Prøvernes betydning og sammenhæng med undervisningen. København: Danmarks Evalueringsinstitut.

Doran, H. C. (2001). Evaluating the Consequential Aspect of Validity on the Arizona Instrument To Measure Standards (No. ED478212). Chicago.

Dorgan, K. (2004). A Year in the Life of an Elementary School: One School's Experiences in Meeting New Mathematics Standards. Teachers College Record, 106(6), 1203-1228.

Ferman, I., Watanabe, Y., & Curtis, A. (2004). The Washback of an EFL National Oral Matriculation Test to Teaching and Learning. In Washback in language testing: Research contexts and methods (pp. 191-210p.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Ferrara, S., & et al. (1988). Local Assessment Responses to a State-Mandated Minimum-Competency Testing Program: Benefits and Drawbacks (No. ED294892). New Orleans.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test Preparation in New Jersey: Inquiry-Oriented and Didactic Responses. Assessment in Education Principles Policy and Practice, 11(1), 67-88.

Fuchs, L. S., & et al. (1989). Monitoring Reading Growth Using Student Recalls: Effects of Two Teacher Feedback Systems. Journal of Educational Research, 83(2), 103-110.

Gayford, C. (1988). Aims, Purposes and Emphasis in Practical Biology at Advanced Level--A Study of Teachers' Attitudes. School Science Review, 69(249), 799-802.

Goldberg, G. L., & Roswell, B. S. (1999). From perception to practice: The impact of teachers' scoring experience on performacne-based instruction and classroom assessment. Educational Assessment, 6(4), 257-290.

*Griffin, P. E., & Smith, P. G. (1997, 1997). The implications of outcome-based education for teachers' work, Brisbane.

Gutkin, J. (1985). The Effect of Diagnostic Inservice Training on Class Reading Achievement and the Number of Lessons Covered. **

Harlen, W., & Crick, R. D. (2003). A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills. London: EPPI.

Harlen, W., & Ruth Deakin, C. (2002). A systematic review of the impact of summative assessment and tests on students motivation for learning. London: EPPI.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007a). Measuring How Benchmark Assessments Affect Student Achievement. Issues & Answers. REL 2007-No. 039.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007b). A Second Follow-Up Year for "Measuring How Benchmark Assessments Affect Student Achievement." REL Technical Brief. REL 2008-No. 002.

Higgins, N., & Rice, E. (1991). Teachers' Perspectives on Competency-Based Testing. Educational Technology, Research and Development, 39(3), 59-69.

Howe, M. E., & Thames, D. G. (1996). Mississippi Reading Teachers' Perceptions toward the Interpretation of Results from Reformed Standardized Assessment in Mississippi. not found(not found), ort.

Izard, J., Jeffrey, P., Silis, G., & Yates, R. (1999). Testing for teaching purposes: application of item response modelling (IRM) teaching focussed assessment practices and the elimination of learning failure in schools. In Learning disabilities: advocacy and action (pp. 163- 187): Parkville Vic: Australian Resource Educators' Association.

Jia, Y., Eslami, Z. R., & Burlbaw, L. M. (2006). ESL Teachers' Perceptions and Factors Influencing Their Use of Classroom-Based Reading Assessment. Bilingual Research Journal, 30(2), 407-430.

Jones, G. W. J. B. R. C. L. Y. T. D. M. (1999). The Impact of High Stakes Testing on Teachers and Students in North Carolina. Phi, Delta, Kappan, 81(3), 199-204.

Kozulin, A., & Garb, E. (2004). Dynamic Assessment of Literacy: English as a Third Language. European Journal of Psychology of Education, 19(1), 65-77.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence From Survey Data and School Performance. Educational Assessment, 8(4), 279-315.

Le Floch, K. C., & et al. (2007). State and local implementation of the no child left behind act vol 3. Washington: U.S. Department of Education.

Luxia, Q. (2004). Has a High-Stakes Test Produced the Intended Changes? In L. W. Cheng, Y.; Curtis, A. (Ed.), Washback in language testing: Research contexts and methods (pp. 171-190). Mahwah, NJ: US: Lawrence Erlbaum Associates Publishers.

Luxia, Q. (2005). Stakeholders' Conflicting Aims Undermine the Washback Function of a High-Stakes Test. Language Testing, 22(2), 142-173.

Luxia, Q. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. Assessment in Education, 14(1), 51-74.

Madelaine, A., & Wheldall, K. (2004). Teachers' reactions to curriculum-based passage reading test data. Special Education Perspectives, 13(1), 55-65.

Malcolm, H. C., Byrne, M., & Harlen, W. (1995). Teachers' assessment and national testing in primary schools in Scotland: roles and relationships. Assessment in Education, 2(2), p129-144.

Mason, R. (2005). The effect of formal assessment on secondary school art and design education. London: EPPI.

Mattsson, H. (1989). Proven i skolan sedda genom lärarnas ögon. Umeå: Pedagogiska inst., Umeå Universitet.

Mildner, T. (1989). A Statewide Assessment and Evaluation of Fourth Grade Mathematics Delivery System. not found(not found), ort.

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a State Performance Assessment Program in Reading and Writing. Educational Research and Evaluation, 12(3), 239-269.

Parker, D. L., & Picard, A. J. (1997). Portraits of Susie: Matching Curriculum, Instruction, and Assessment. Teaching Children Mathematics, 3(7), 376-382.

Pinder, P. J. (2008). A Critique Analysis of NCLB, Increase Testing, and Past Maryland Mathematics and Science HSA Exams: What Are Maryland Practitioners' Perspectives?

Prueher, J. (1987). Improving Achievement and Attitudes in Elementary Algebra through Written Error-Correcting Feedback and Free Comments on Tests. not found(not found), ort.

Roach, A. T., Elliott, S. N., & Berndt, S. (2007). Teacher perceptions and the consequential validity of an alternate assessment for students with significant cognitive disabilities. Journal of Disability Policy Studies, 18(3), 168-175.

Ryan, J., & Williams, J. (2000). National testing and the improvement of classroom teaching: can they coexist? British Educational Research Journal, 26(1), p49-73.

Ryan, P. (1994). Teacher Perspectives of the Impact and Validity of the Mt. Diablo Third Grade-Curriculum-Based Alternative Assessment of Mathematics (CBAAM). not found(not found), ort.**

Scott, C. (2007). Stakeholder perceptions of test impact. Policy & Practice, 14(1), 27-49.

Shannon, A. J. (1980). Effects of Methods of Standardized Reading Achievement Test Administration on Attitude toward Reading. Journal of Reading, 23(8), 684-686.

Shohamy, E., & et al. (1996). Test Impact Revisited: Washback Effect Over Time. Language Testing, 13(3), 298-317.

Silis, G. (2005a). How a school has used data to support learning. In J. B. Mousley, L. & Campbell, C. (Ed.), Mathematics: celebrating achievement, 100 years, 1906-2006 (pp. 287-297). Brunswick, Victoria: Mathematical Association of Victoria.

Silis, G. (2005b, 2005). Using test information to improve program delivery.

Smart, M. (2004). An investigation into the consequential validity of the Secondary Entrance Assessment examination. Florida State University, Tallahassee.

Smith, P. S., & et al. (1992). The Impact of End-of-Course Testing on Curriculum and Instruction. Science Education, 76(5), 523-530.

Stecher, B., Chun, T., & Barron, S. (2004). The Effects of Assessment-Driven Reform on the Teaching of Writing in Washington State.MT: Monograph Title. In L. W. Cheng, Y.; Curtis, A. (Ed.), Washback in language testing: Research contexts and methods (pp. 237). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. Applied Measurement in Education, 16(1), 1-26.

Study, N. F. S. (1992). The Influence of Testing on Teaching Math and Science in Grades 4-12. Chestnutt Hill, MA: NFS Study.

Stullich, S., Eisner, E., & McCrary, J. (2007). National Assessment of Title I. Final Report. Volume I: Implementation. NCEE 2008-4012.

Sturman, L. (2003). Teaching to the test: science or intuition? Educational Research, 45(3), p261-273.

Tresch, S. (2007). Potential Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen. Bern: H.E.P. Verlag.

Viera, D. R. (1986). Remediating Reading Problems in a Hispanic Learning Disabled Child from a Psycholinguistic Perspective: A Case Study. Journal of Reading, Writing, and Learning Disabilities International, 2(1), 85-97.

Wall, D. (1996). Introducing New Tests into Traditional Systems: Insights from General Education and from Innovation Theory. Language Testing, 13(3), 334-354.

Wall, D., & Alderson, J. C. (1992). Examining Washback: The Sri Lankan Impact Study. not found(not found), ort.

Watanabe, Y. (2004). Teacher Factors Mediating Washback. In L. W. Cheng, Y.; Curtis, A. (Ed.), Washback in language testing: Research contexts and methods (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Winfield, L. F. (1987). The relationship between minimum competency testing programs and students' reading proficiency. Implications from the 1983-84 National Assessment of Educational progress in reading and writing. (No. ED283841).

Winter, J., Fitz, J., & Firestone William, A. (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. Assessment in Education, 7(1), p13-37.

10 Referencer

Her opregnes alene referencer, der indgår i det systematiske reviews kommentartekst. Den samlede oversigt over de i reviewet indgående undersøgelser er gengivet i Kapitel 9.

Allerup, P. P. (1987). *Raschmodeller - principper og anvendelse*. København: Danmarks pædagogiske Institut, nr. 29.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

De nationale test. (Verificeret 2009.03.28). from http://evaluering.uvm.dk/templates/laerereOgLedere_layout.jsf?state=LAERERE_OG_LEDERE:true:true:true:true:true:false:false:false:false:%5Bnull%5D:COWI:LAERERE_NAT_TEST:8:

Depart. f. Children, Schools, & Families. (Verificeret 2008.10.14, 28 April 2009). Major reforms to school accountability including an end to compulsory national tests for fourteen year olds. from http://www.dcsf.gov.uk/pns/DisplayPN.cgi?pn_id=2008_0229

Duit, R. (2009). *Bibliography of Students' and Teachers' Conceptions and Science Education*. from http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html

Institute for Objective Measurement. from http://www.rasch.org/

Jussim, L., & Harber, K. (2005). Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and social psychology review, 9*(2), 131.

Kavli, H. (2008). *Nasjonale prøver 2007 - Brukernes evaluering av gjennomføringen*. Oslo: Synovateo. Document Number)

Kavli, H., Kalve, A., & Tamsfoss, S. (2005). *Analyserapport - Evaluering av gjennomføringen av de nasjonale prøver*. Oslo: MMI.

Korp, H. H. (2003). *Kunskapsbedömning - hur, vad och varför*. Kalmar: Myndigheten för skoleutveckling.

Lie, S., Caspersen, M. L., & Björnsson, J. K. (2004). *Nasjonale prøver på prøve. Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2004*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Lie, S., Hopfenbeck, T. N., Ibsen, E., & Turmo, A. (2005). *Nasjonale prøver på ny prøve. Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Mejding, J. (1994). *Den grimme ælling og svanerne? om danske elevers læsefærdigheder* København: Danmarks Pædagogiske Institut.

Nordenbo, S. E. (2008). Fra progressiv til liberal pædagogik. In F. Collin & J. Faye (Eds.), *Ideer vi lever på. Humanistisk viden i videnssamfundet* (pp. 93-110). København: Akademisk Forlag.

Nordenbo, S. E., Jensen, B., Johansson, I., Kampmann, J., Søgaard Larsen, M., Moser, T., et al. (2008). *Forskningskortlægning og forskervurdering af skandinavisk forskning i året 2006 i institutioner for de 0-6 årige*. København: Danmarks Pædagogiske Universitetsforlag og Dansk Clearinghouse for Uddannelsesforskning.

Nordenbo, S. E., Jensen, B., Johansson, I., Kampmann, J., Søgaard Larsen, M., Moser, T., et al. (2009). *Forskningskortlægning og forskervurdering af skandinavisk forskning i året 2007 i institutioner for de 0-6 årige*. København: Danmarks Pædagogiske Universitetsforlag og Dansk Clearinghouse for Uddannelsesforskning.

Nordenbo, S. E., Søgaard Larsen, M., Tiftikçi, N., Wendt, R. E., & Østergaard, S. (2008). *Lærerkompetanser og elevers læring i barnehage og skole: et systematisk review utført for Kunnskapsdepartementet, Oslo* (3. ed.). København: Danmarks Pædagogiske Universitetsforlag: Dansk Clearinghouse for Uddannelsesforskning.

NOU. (2002). *Førsteklasses fra første klasse. Forslag til rammeverk for et nasjonalt kvalitetsvurderingssystem av norsk grunnopplæring*. Oslo: Statens forvaltningstjenesteo. Document Number)
NOU. (2003). *I første rekke. Forsterket kvalitet i en grunnopplæring for alle*. Oslo: Statens forvaltningstjenesteo. Document Number)
Obligatoriske test i folkeskolen. (2006). from http://web.archive.org/web/20070518105136/http://www.uvm.dk/evalueringskultur/test/adaptivetest.htm
OECD. (1989). *OECD-vurdering av norsk utdanningspolitikk. Norsk rapport til OECD. Ekspertvurdering fra OECD*. Oslo: Kirke- og undervisningsdepartementet. Kultur- og vitenskapsdepartementet. Organisasjonen for kulturelt og økonomisk samarbeid (OECD). Aschehoug.
Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology and Community Health, 57*, 527-529.
Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences. A Practical Guide*. Malden: Blackwell Publishing.
Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., et al. (2006). *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. Version 1, April 2006*: A Product from the ESRC Methods Programme.
Rieper, O., & Foss Hansen, H. (2007). *Metodedebatten om evidens*. København: AKF Forlaget.
Rosenthal, R., & Jacobson, L. (1977). *Pygmalion i klasseværelset*. København: Gyldendal.
Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom. Expanded edition*. New York:: Irvington.
TNS-Gallup. (2004). *Rektorers og læreres erfaringer med de nasjonale prøvene 2004*. from http://www.utdanningsdirektoratet.no/upload/Rapporter/Evaluering_av_de_nasjonale_provene_2004.pdf
UFD. (2002). *St.prp. nr. 1 (2002-2003)*. Oslo: Utdannings- og forskningsdepartementeto. Document Number)
UFD. (2004). *St.meld. nr. 30 (2003-2004): Kultur for læring*.
Weiss, C. H. (1998). *Evaluation. Second Edition*. Upper Saddle River, New Jersey: Printice Hall.
Wholey, J. S. (1987). *Organizational excellence: Stimulating quality and communicating value*. Lexington, MA: Lexington Books.