Test, og hvad enhver læseunderviser bør vide om dem

MADS POULSEN, LEKTOR, CENTER FOR LÆSEFORSKNING, KØBENHAVNS UNIVERSITET
OG CARSTEN ELBRO, PROFESSOR, CENTER FOR LÆSEFORSKNING, KØBENHAVNS UNIVERSITET

Faglige test kan være fine iagttagelsesredskaber, som kan supplere lærerens faglige indtryk af eleven. Som alle andre redskaber kræver test en vis faglig indsigt i deres brug. Denne artikel giver en introduktion til formål med testning og kvalitetskrav til test og testning. En vigtig pointe er, at man bør stille visse kvalitetskrav til formelle test, fx at de har rimelig og velbeskrevet gyldighed og pålidelighed, men at man ikke kan forvente, at test er så informative, at de overflødiggør kvalificererede fagpersoner.

De fleste af os trives og udvikler os bedst, hvis vi oplever udfordringer, som på den ene side er overkommelige og på den anden side er tilpas udfordrende til at holde vores opmærksomhed. Børn – som voksne – er sammensatte individer. Det kan gøre det vanskeligt at afgøre, hvad der er passende udfordringer. Ida kan være virkelig god til matematik, men synes, at dansk er kedeligt. Muhammed er måske god til at stavning, afkodning og historie, men har vanskeligt ved at forstå tekster, fordi han kender færre ord på dansk end mange af sine jævnaldrende.

Gad vide hvorfor Ida synes, at dansk er kedeligt? Det kan der være så mange grunde til. Man kan både synes, at noget er kedeligt, hvis det er for let eller for svært. Måske har hun faktisk lidt svært ved at læse. Men hun er jo god til matematik, og hun har sådan nogle fornuftige forældre, og så kan man måske som travl lærer hurtigt komme væk fra at tænke mere over det. Men hverken matematikfærdigheder eller pæne forældre sikrer strengt taget børn mod læsevanskeligheder. Det gør ikke sagen lettere for læreren, at den enkelte elevs situation kan ændre sig. En ændring i Idas hjemmesituation kan gøre, at hun i en længere periode ikke suger så meget til sig og pludselig faktisk har svært ved matematik. Og måske har Muhammed brugt sine gode afkodningsfærdigheder til at læse en masse og er derigennem i gang med at opbygge et glimrende ordforråd.

Det er meget at forlange, at lærere skal kunne holde styr på 25 komplicerede elever i flere klasser alene ved hjælp af uformelle iagttagelser og intuition. Der er brug for faglige iagttagelsesredskaber. Uformelle iagttagelser er en vigtig og naturlig del af undervisningen: Hvordan løser eleverne de opgaver, som bliver stillet i timerne? Men øvelserne i timerne er designet med henblik på at være lærerige for så mange som muligt, ikke på at afdække elevfærdigheder. Og aktiviteterne gennemføres gerne i et virvar af gruppearbejde og andre aktiviteter, som gør det svært at vurdere, hvad de enkelte elever egentlig foretager sig. Derfor kan det være fornuftigt at supplere med test.

Hvad er test?

En test er bare en veldefineret måde at iagttage elevers viden og færdigheder på. Testen består typisk af nogle opgaver eller spørgsmål, som stilles

under de samme betingelser til alle, og som således stiller alle elever lige. De fælles betingelser betyder også, at testen er den samme over tid og derfor kan bruges til at følge elevers udvikling med. Et testresultat er objektivt i den forstand, at det ikke er farvet af lærerens øvrige viden om eleven. Sådanne objektive informationer kan supplere, men ikke erstatte lærerens helhedsindtryk af eleven.

En veldefineret måde at iagttage viden og færdigheder på er ikke noget, man kan have så meget imod. Det er derimod *brugen af testresultatet*, der kan være problematisk

En veldefineret måde at iagttage viden og færdigheder på er ikke noget, man kan have så meget imod. Det er derimod *brugen af testresultatet*, der kan være problematisk. *Ingen* bryder sig om test, der *alene* fører til stempling af en elev som dum eller uduelig. For eksempel er en typisk IQ-test bare nogle opgaver af samme uskyldige slags, som man kan more sig med en regnvejrsdag i sommerferien. Opgaverne – testen i sig selv – kan næppe skade nogen. Det er *brugen af resultatet*, der kan være ødelæggende fx for elevens selvopfattelse eller status i andres øjne.

Status og fremadrettet brug

Derfor er det forholdsvis almindeligt i teoretiske uddannelsessammenhænge at skelne mellem (formativ, fremadrettet) brug af test til at tilrettelægge undervisning ud fra og (summativ, status) brug af test til at vurdere elevens udbytte af undervisningen.

I praksis er det ikke altid let at skelne, fordi den *samme* test kan *bruges* på begge måder. Et resultat i den nationale læsetest i 4. klasse kan i princippet bruges (formativt) til at tilrettelægge den kommende læseundervisning ud fra og til (summativt) at vurdere elevernes udbytte af den foregående undervisning.

Hvis man løfter blikket fra den enkelte elev til *underviserens* perspektiv, så kan en summativ udbyttetest faktisk i høj grad blive en formativ tilrettelæggelsestest. For underviseren har selvfølgelig brug for at vide, om undervisningen virkede efter hensigten – at eleven faktisk lærte sig det, der var sigtet med undervisningen. Det kan muligvis være for sent at planlægge noget for den aktuelle elev, fx fordi undervisningen er slut; men *underviseren* kan tage ved lære af elevudbyttet og søge efter veje til at forbedre undervisningstilbuddet. Så i den forstand kan underviseren bruge testresultatet fremadrettet ved planlægning af undervisning for kommende elever (se eksempler i Elbro & Poulsen, 2015).

Testformater

En test af læseforståelse kan se ud på mange måder. Den prototypiske test indeholder en eller flere tekster og en række spørgsmål til dem, som eleven skal besvare. Svarmulighederne kan være givet (multiple choice), og eleverne skal så vælge det rigtige svar. Eller eleven skal selv formulere svaret. Men der er mange andre formater. Et andet format er clozetest, hvor ord er taget ud af teksten, og læseren så skal forsøge at regne de manglende ord ud.

Nogle test ligner noget fra virkeligheden mere end andre test. For den uerfarne kan det fx være betryggende at vide, at en læsetest indeholder autentiske tekster og spørgsmål. Det svarer til, at mennesker, der ikke har set et termometer før, heller ikke har så megen tillid til det og hellere selv vil føle efter, om der er tilpas koldt eller for varmt i køleskabet. Man skal lige vænne sig til testformater, der som fx clozetest ikke ligner noget fra virkeligheden; men derefter kan de være helt fine og give præcise og brugbare oplysninger på ret kort tid.

Adaptive test. Langt de fleste test indeholder de samme opgaver til alle deltagere. Men der er også test, som tilpasser sig elevens færdigheder, såkaldt adaptive test, som fx de nationale test i læsning. Ved adaptive test flytter dygtige elever hurtigt frem til nogle af de sværere opgaver. Og testningen standser, når opgaverne bliver klart for svære for eleven. Det format har den tekniske fordel, at der

kan blive flere relevante opgaver på elevens niveau til at give et godt og sikkert billede af elevens færdigheder. Desuden vil alle elever opleve testen som ca. lige svær, fordi der er opgaver, som selv de dygtigste elever ikke kan løse. Om det er en fordel, at alle finder testen lige svær, kan faktisk diskuteres, fordi elever også kan have nytte af at få korrigeret en overvurderet eller undervurderet selvopfattelse.

Gruppetest eller individuel undersøgelse?
Inden man går i gang med en individuel undersøgelse med fx den nationale ordblindetest, så skal der være gode grunde. Som i Idas tilfælde ovenfor kan man have en række uformelle iagttagelser fra undervisningen, som kunne tyde på, at Ida har vanskeligheder. Normalt vil man supplere sådanne iagttagelser med resultaterne af en gruppetest, dvs. en test, som hele klassen (eller en større gruppe) har gennemført på samme tid (national test eller forlagsfremstillet test). Gruppetest kan netop bruges til at screene for eventuelle vanskeligheder. Hvis Ida så viser sig at have vanskeligheder med gruppetesten, bør man gå videre med en individuel udredning.

Præcision, hastighed og effektivitet

Når resultaterne af en videns- eller færdighedstest så skal gøres op, er der typisk de følgende tre samlede resultater:

Hastighed: Hvor mange opgaver nåede eleven? Det opgøres typisk som antal nåede opgaver (eller ord) pr. minut. Man tæller alle de nåede opgaver, uanset om de er besvaret rigtigt, forkert eller opgivet.

Præcision: Hvor mange af de nåede opgaver havde eleven rigtig? Det opgøres normalt i procent.

Samlet færdighed eller effektivitet: Antal rigtigt besvarede opgaver pr. minut. Det er hastigheden ganget med præcisionen.

Da skriftsproglige færdigheder netop er færdigheder, er både hastighed og præcision væsentlige aspekter af færdigheden. En læsning med blot 30 ord i minuttet giver en elev i 8. klasse klare praktiske udfordringer, selv om den måtte være 100 % korrekt.

Over eller under grænsen?

Skriftsproglige test bruges mange gange til at finde de elever, som har brug for særlig opmærksomhed og evt. støtte. Nogle gange er test direkte indrettet med det formål for øje, som det fx er tilfældet med den nationale ordblindetest. Testen må så have en eller flere vejledende grænser for, hvornår et resultat er bekymrende, og hvornår det ikke er. Disse grænser bør være veldokumenterede (se om gyldighed nedenfor).

Kvaliteten af test

Den pædagogiske nytte af en faglig test afhænger af 1) om testen er passende til formålet, 2) at testen gyldigt måler det, den skal måle, 3) at testen er pålidelig, 4) at der følger et sammenligningsgrundlag med, så man kan vurdere resultaterne, og 5) at testen bliver afviklet og fortolket fornuftigt af kompetent fagpersonale, som samtidigt har blik for hele eleven.

1. Passende til formålet. Valg af test bestemmes af formålet med at teste. Læreren kan fx have en fornemmelse af, at Ida har svært ved at læse tekster med god forståelse. Derfor er det relevant at se på Idas resultater med en gruppetest af læseforståelse (= tekstforståelse i de nationale test eller i en forlagsproduceret test). Hvis lærerens fornemmelse bliver bekræftet af testen, er det oplagt at gå videre for at finde ud af, hvor Idas vanskeligheder opstår i læsningen. De kan jo både have med ordafkodningen og med sprogforståelsen at gøre (se fx Elbro, 2014, kapitel 2). Måske er der allerede screeningsresultater, som kan kaste lys over Idas færdigheder. Måske skal der en nærmere individuel undersøgelse til. Ida kan have svært ved det ene uden at have svært ved det andet.

Uanset hvad, er målet med undersøgelserne (og testningen) hele tiden at finde ud af, *hvilke* opgaver og aktiviteter der er relevante for Ida, og hvilket niveau de skal være på. Ida kan have brug for lette tekster og opgaver i afkodning af ord. Ida kan også have brug for at lære at trække mere på sin baggrundsviden for at forstå tekster. Man kan ikke vide det uden at vide, hvor Ida har sine stærke og svage sider.

2. Gyldighed (validitet). En test skal faktisk og primært teste det, den påstår at teste. En test, hvor opgaverne består i, at man skal benævne billeder, er helt banalt ikke en test af afkodningsfærdighed. En sådan test udfordrer faktisk ikke afkodningsfærdigheder. En test, hvor opgaverne består i, at man skal læse en kompliceret faglig tekst og besvare spørgsmål til teksten, er nok ikke primært en test af afkodningsfærdighed. Den udfordrer måske faktisk afkodningsfærdighed. Men den udfordrer også andre færdigheder, fx ordforråd, baggrundsviden og kompliceret sprogforståelse, så den er ikke primært en test af afkodningsfærdighed. Problemet er, at hvis en elev svarer forkert, så ved man ikke, om det skyldes manglende afkodningsfærdighed, ordforråd, baggrundsviden eller kompliceret sprogforståelse. Der er snarere tale om en generel læseforståelsestest.

Vurderingen af gyldighed starter således med en faglig vurdering af, hvad der er udfordrende i testen. Men selv eksperter kan have svært ved fra skrivebordet at vurdere, hvad der i nogle bestemte opgaver egentlig er udfordrende for en bestemt målgruppe.

Derfor kan det være relevant at undersøge, om vurderingen af gyldighed stemmer med systematiske observationer fra virkeligheden. Resultaterne fra en ny læseforståelsestest skulle fx gerne stemme overens med en etableret test af læseforståelse. De skulle gerne have samstemmende gyldighed. Dvs. at hvis man tester et større antal elever, så skulle de elever, der scorer højt på den ene test, også helst score højt på den anden, og dem, der scorer lavt på den ene, skulle gerne score lavt på den anden. Ellers giver det ikke mening at sige, at de to test er gyldige test af samme færdighed. En tommelfingerregel er, at sammenhængen mellem de to test bør være stærk (en korrelation over 0,7; se fx Elbro & Poulsen, 2015, kapitel 5 om korrelation).

Nogle test er konstrueret og bruges med det formål at vurdere, om man skal være så bekymret for et barns faglige udviklingsmuligheder, at man skal lave en særlig indsats for barnet. Når testen skal forudsige noget, så vil man forvente, at der er dokumentation for, at den faktisk kan forudsige, sådan som det påstås. Man forventer *forudsigende* (*prædiktiv*) gyldighed af testen. Forudsigende gyldighed er fx relevant for sprogvurderingstest i dagtilbuddene. Hvis de børn, der scorer lavt i sprogvurderingsmaterialet, også har særligt stor tendens til senere at have sproglige vanskeligheder, så giver sprogvurderingsmaterialet gyldig information. En grov tommelfingerregel er, at forudsigelsessikkerheden skal være over 80 %, hvis den skal være pædagogisk brugbar for de enkelte børn. Sikkerheden skal gælde både for børn, der udpeges til at være i risikozonen, og for børn, der ikke udpeges til at være i risikozonen. Hvis forudsigelsen strækker sig længere ud i fremtiden end ca. et år, så bliver den selvfølgelig svagere – ligesom vejrudsigten bliver mere usikker, jo længere ud i fremtiden den strækker sig. Man kan læse mere om denne slags analyser og vurderinger i Poulsen (2018) og se det eksemplificeret i en artikel om Ordblinderisikotesten (Gellert & Elbro, 2018).

3. Pålidelighed (reliabilitet) og følsomhed. En test skal give et rimeligt pålideligt indtryk af elevens færdighed. Det duer ikke, hvis den samme elev får vurderet sin færdighed som høj den ene uge, men lav den næste uge. Så kan resultatet ikke bruges til ret meget (forudsat at det giver mening at bruge den samme test to uger i træk). Testresultater bliver upålidelige, hvis tilfældigheder får lov til at spille for stor en rolle.

Blandt andet derfor skal en test helst have mange opgaver, for ellers kan små tilfældigheder i enkeltbesvarelser spille en stor rolle for testresultatet. Hvis der fx kun var ét ord i en ordforrådstest, fx *mikroskop*, så ville det være ret tilfældigt, hvem der klarede sig godt. Elever, der lige havde arbejdet med mikroskoper i natur og teknik, ville klare sig godt – uanset hvilke ord de ellers kendte. Sådanne tilfældigheder kan aldrig fjernes. Men hvis der er tilstrækkeligt mange og velvalgte enkeltopgaver, så vil en enkelt lille tilfældighed her og der ikke spille den store rolle for det samlede resultat. Mange enkeltopgaver er også en forudsætning for, at testen overhovedet kan skelne mellem mange niveauer i færdigheder, dvs. at testen har en god følsomhed.

Det nødvendige antal enkeltopgaver i en test afhænger af mange ting, bl.a. hvor følsom testen skal være, og hvor pålidelig hver enkeltopgave er. Hvis det fx er forholdsvis let at gætte det rigtige svar (fordi der blot er to eller tre svarmuligheder at

vælge imellem), så er følsomheden lav, og der skal flere enkeltopgaver til.

Man kan vurdere pålideligheden af en test ved at se på faktiske testresultater. Man kan fx lade den samme elevgruppe tage den samme test (eller en paralleludgave) med kort mellemrum. Så vurderer man *gentest-pålideligheden*. Eller man kan se, hvordan resultaterne med den ene halvdel af enkeltopgaverne hænger sammen med resultaterne af den anden. De forskellige metoder bygger alle på en idé om, at en tests pålidelighed kan måles som testens korrelation med sig selv. En tommelfinderregel er, at pålideligheden skal være over 0,7. Du kan læse mere om dette i Elbro & Poulsen (2015, kapitel 7). En forlagsfremstillet test bør indeholde oplysninger om testens pålidelighed. Hvis en lærer selv fremstiller en lille test til uformelt at vurdere, hvad eleverne kan, så er der sjældent grund til at besvære sig med en egentlig pålidelighedsundersøgelse. Men man bør så være opmærksom på, at der er grænser for, hvor meget man skal lægge i resultaterne. Hvis man vil have præcis information, er det som oftest bedre at vælge et testredskab med dokumenteret gyldighed og pålidelighed.

Ingen test er 100 % pålidelig. Mindre kan også gøre det, især hvis man forstår at sammenholde testresultater med hinanden og med mere uformelle iagttagelser af eleverne.

4. Sammenligningsgrundlag (normer). Det er let at opgøre en elevs testscore, fx til 83 % rigtige i en diktat. Men sådan et tal er sjældent informativt i sig selv. Er 83 % godt eller dårligt? Det kan komme an på så meget, fx ordene i diktaten, og om eleven gik i anden eller femte klasse. Det kan være svært ud af det blå at sige, hvad man egentlig kan forvente af elever på et givet klassetrin. Selv erfarne lærere kan blive overraskede over, hvad der har sat sig fast efter et undervisningsforløb. Man har brug for et sammenligningsgrundlag. Derfor følger der med mange forlagstest oplysninger fx om gennemsnitsscorer for bestemte klassetrin, eller hvor stor en andel af eleverne der fordeler sig i forskellige resultatgrupper. Sådanne normer er typisk baseret på resultater fra mange elever spredt på flere skoler. De fungerer nærmest som mange års systematisk forarbejdede erfaringer for en enkelt lærer.

Testresultater bør altid sammenholdes med andre iagttagelser, når hensigten er at tilrettelægge undervisningen for den enkelte elev.

Hvis Ida i midten af tredje klasse scorer væsentligt under andre elever i tredje klasse i en afkodningstest, så er der grund til at være opmærksom på, om hun får tilstrækkeligt ud af de læseafhængige opgaver, som man synes passer til de øvrige elever i klassen. Og måske skulle man sørge for, at hun får noget mere træning med lettere afkodningsopgaver, som man ellers ville mene passede til yngre elever.

5. Kvalificeret afvikling og fortolkning. Det er vigtigt, at man afvikler test på den måde, som er beskrevet i testvejledningen. Det er en forudsætning for, at man kan sammenligne testscorerne med normerne. Hvis ikke eleverne løser opgaverne under samme vilkår, så er sammenligning umulig. Det er også vigtigt, at opgaverne og instruktionerne er klare, så eleverne har samme forståelse af, hvad de skal.

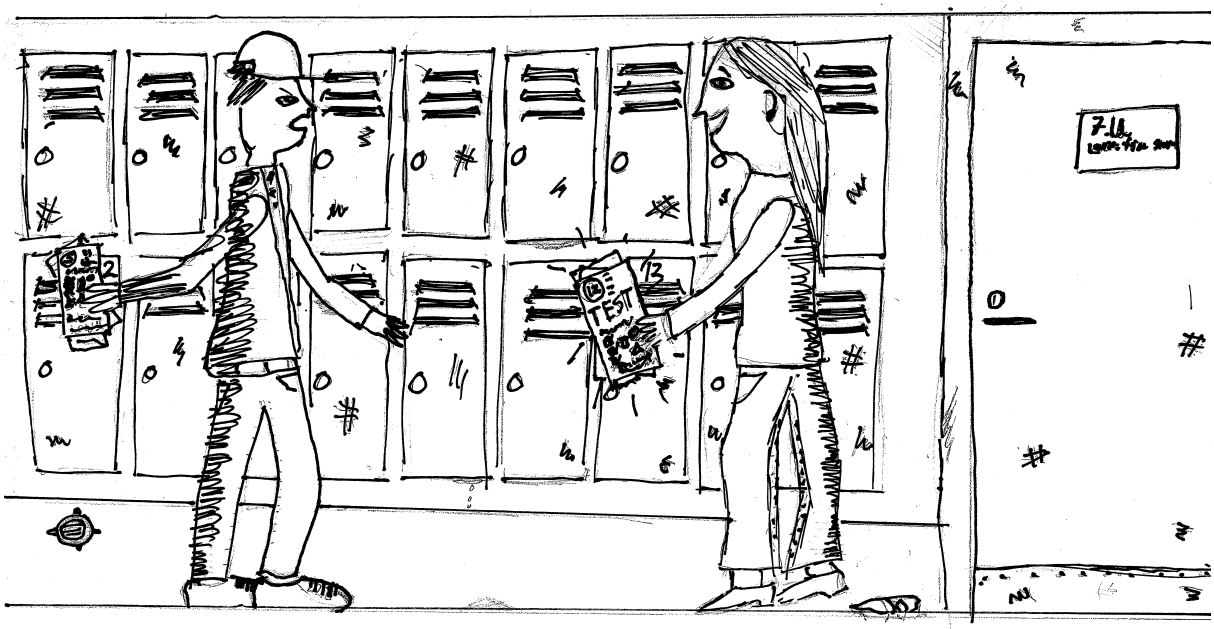
Endelig er det vigtigt, at test udvælges og fortolkes af fagfolk med de nødvendige kvalifikationer og helst med et bredt kendskab til eleven. Faglig indsigt er nødvendig for at vælge, hvilke færdigheder eller delfærdigheder testen skal afdække, og hvilke opfølgende tiltag der er relevante at iværksætte. Fx skal man have læsefaglig indsigt i forskellen mellem afkodning og sprogforståelse for at udvælge og følge op på relevante test af delfærdigheder i læsning. Man skal også have en smule testfaglig indsigt for fornuftigt at kunne se en testscore i forhold til en norm. Lidt testfaglig indsigt hjælper også en til at forholde sig konstruktivt til, at test ikke er 100 % pålidelige. Testresultater bør altid sammenholdes med andre iagttagelser, når hensigten er at tilrettelægge undervisningen for den enkelte elev.

Referencer

Elbro, C., & Poulsen, M. (2015). *Hold i virkeligheden. Statistik og evidens i uddannelse.* København: Hans Reitzels Forlag.

Gellert, A. S., & Elbro, C. (2018). Forudsigelse af alvorlige afkodningsvanskeligheder i 2. klasse på basis af testresultater i 0. og 1. klasse. *Pædagogisk Psykologisk Tidsskrift, 55,* 90-103.

Poulsen, M. (2018). The challenge of early identification of later reading difficulties. *Perspectives on Language and Literacy, 33,* 11-14.



Tegning: Jakob Martin Majholm