Eindrapport overzichtsstudie

Ontwerprichtlijnen voor formatief toetsen vanuit de geheugenpsychologie

1 + 1 = 3?

Subsidienummer: 405-17-711

Kim Dirkx, Desirée Joosten-ten Brinke, en Gino Camp
Welten Instituut, Open Universiteit, Heerlen

Postbus 6400, 6401 DL Heerlen

Deze uitgave is mogelijk gemaakt door een subsidie van NRO, het Nationaal Regieorgaan Onderwijsonderzoek (projectnummer: 405-17-711). Gebruik en overname van teksten, ideeën en resultaten uit deze publicatie is vrijelijk toegestaan, mits met bronvermelding.

Inhoudsopgave

Samenvatting 4
1. Inleiding 6
1.1. Een brede en smalle definitie van formatief toetsen 7
1.2. Retrieval practice 9
1.3. Distributed practice 14
1.4. Ontwerprichtlijnen voor formatief toetsen 16
1.5. Onderzoeksvraag 17
2. Methode 19
2.1. Literatuursearch en selectiecriteria 19
2.2. Codeerschema 19
2.3. Data-analyse 21
3. Resultaten 22
3.1. Algemeen 22
3.2. De richtlijnen 24
4. Conclusies en discussie 29
Referenties 35

Samenvatting

Hoewel tussentijdse, formatieve toetsen in het onderwijs vaak gebruikt worden om het leerproces te ondersteunen, is er nog weinig wetenschappelijke onderbouwing voor de wijze waarop dit soort toetsen ontworpen zouden moeten worden om de grootste effecten te sorteren. In de geheugenpsychologie wordt al decennia lang onderzoek gedaan naar de effecten van *retrieval practice op leren*, ofwel het oefenen met het tussentijds ophalen van informatie uit het geheugen. De procedure die hiervoor wordt gebruikt vertoont grote overeenkomsten met het proces van formatieve toetsing, waar ook leerstof tussentijds formatief wordt getoetst ter voorbereiding op een summatieve eindtoets. Hoewel het onderzoek naar *retrieval practice* in eerste instantie voornamelijk in het psychologisch laboratorium plaatsvond, is er de laatste vijftien jaar meer en meer onderzoek gedaan in de onderwijspraktijk. Dit onderzoek resulteert in richtlijnen voor het ontwerpen van tussentijdse toetsen met een optimaal effect op leren. Het is bijvoorbeeld bekend dat het soort toets en ook de spreiding tussen opeenvolgende formatieve toetsen het leren van tussentijds toetsen kan beïnvloeden.

Echter, in de literatuur over formatieve toetsing, wordt nog nauwelijks gebruik gemaakt van de principes die beschreven worden in de geheugenpsychologie. Middels een systematische literatuurstudie hebben we geprobeerd de twee onderzoekslijnen rondom formatief toetsen en geheugenpsychologie bij elkaar te brengen en *evidence-based* richtlijnen te formuleren voor het ontwerpen van formatieve toetsen in de onderwijspraktijk. De literatuurstudie is uitgevoerd in het voorjaar van 2018 en leverde uiteindelijk 110 artikelen op die aan de hand van vijf ontwerpvragen (van der Vleuten & Driessen, 2000) geanalyseerd zijn.

Clustering van de resultaten leverde uiteindelijk 10 wetenschappelijk onderbouwde richtlijnen op voor het ontwerpen van formatieve toetsen voor het onderwijs:

1. Gebruik formatieve toetsen in verschillende domeinen en bij verschillende soorten leermaterialen (bijv. teksten, woordjes, sommen, presentaties en video's) om leren te stimuleren;

2. Gebruik formatieve toetsen in elk geval voor onthouden, begrijpen, en toepassen van informatie;
3. Stem het niveau en de inhoud van de formatieve toets af op de eindtoets;
4. Kies voor een combinatie van open- en gesloten vragen bij formatieve toetsen;
5. Als je formatief toetst, zorg dan dat je in de feedback het goede antwoord geeft;
6. Zet een formatieve toets pas in na een initiële leerfase;
7. Toets dezelfde stof minstens één keer maar maximaal drie keer;
8. Spreid de toetsen uit over de tijd;
9. Begin niet vlak voor de summatieve toets met het maken van formatieve toetsen maar gebruik de 20% regel;
10. Bed formatieve toetsen bewust in het toetsprogramma in, waarbij de programmering geen vrijblijvend maar sturend karakter heeft.

In de resultatensectie van dit rapport worden de richtlijnen toegelicht en in de discussie wordt ingegaan op de beperkingen van deze overzichtsstudie en implicaties van de richtlijnen voor de praktijk.

1. Inleiding

Toetsing is een onlosmakelijk onderdeel van het onderwijs, en wordt traditioneel gezien als hét middel voor kwaliteitsborging, certificering en selectie. Maar toetsen kunnen ook worden ingezet als instrument om het leerproces te ondersteunen en lerenden te informeren over hun ontwikkeling. Wanneer toetsen op deze manier in het onderwijs ingezet worden, spreken we over formatief toetsen of in het Engels *formative assessment*. Door het maken van een formatieve toets krijgt een leerling of student ook inzicht in de vorm en inhoud van een toets, wat bijdraagt aan de transparantie van toetsing. Daarnaast kan een formatieve toets zorgen voor een beter begrip van de leerstof.

Er zijn echter nog veel vragen rondom de voorwaarden waaronder formatieve toetsen een positieve werking hebben op het leren en hoe je formatieve toetsen het meest effectief kunt inzetten. Concrete vragen waar op dit moment geen eenduidig antwoord op te geven is, zijn bijvoorbeeld: Uit wat voor type vragen zou een formatieve toets moeten bestaan? Hoe vaak zou je een formatieve toets moeten maken en hoeveel tijd moet er tussen de toetsen zijn? Deze vragen kunnen vanuit verschillende perspectieven beantwoord worden en tot op heden zijn de antwoorden op deze vragen vanuit de geheugen psychologische literatuur nog niet systematisch bekeken

Echter, de geheugenpsychologie worden effectieve leerstrategieën beschreven die zeer duidelijke overeenkomsten hebben met het proces van formatief toetsen, maar de koppeling tussen leerstrategieën vanuit de geheugenpsychologie en formatief toetsen is zelden gemaakt of heeft nog niet geleid tot concrete richtlijnen. Dit is een witte vlek in het onderwijsonderzoek rondom formatief toetsen, terwijl de verbinding concrete adviezen kan opleveren voor het ontwerp en gebruik van formatieve toetsen in de onderwijspraktijk. Met deze literatuurstudie willen we dan ook concrete richtlijnen formuleren voor het ontwerpen van formatief toetsen vanuit de geheugenpsychologie en bijdragen aan een sterkere theoretische onderbouwing voor het ontwerp en de effecten van formatief toetsen.

1.1. Een brede en smalle definitie van formatief toetsen

Formatief toetsen is geen eenduidig gedefinieerd begrip (zie Sluijsmans, et al., 2013). Het varieert van een brede definitie als leer- en instructiestrategie met een verscheidenheid aan activiteiten om het leren van leerlingen of studenten te stimuleren, tot de smalle definitie van een daadwerkelijke fysieke toets met open of gesloten vragen. In Figuur 1 (uit: Brookhart, 2007, p. 44, en overgenomen uit Sluijsmans et al., 2013) is de ontwikkeling van het concept *formatief toetsen* weergegeven over de jaren heen. In de figuur is te zien dat in de oorspronkelijke definitie het verzamelen van informatie over het leerproces centraal stond (Scriven, 1967). In de loop van de tijd zijn aan deze functie van toetsen verschillende functies toegevoegd. Bloom, Hastings, en Madaus (1971) voegden bijvoorbeeld het gebruiken van de informatie voor onderwijs-beslissingen toe. En Sadler (1989) vulde de definitie van formatief toetsen aan met de notie dat de verzamelde informatie ook door de lerenden gebruikt zou moeten worden om het eigen leerproces te bevorderen. De meest recente concepties houden aan dat formatief toetsen niet alleen informatie op moet leveren die gebruikt kan worden voor het verder vormgeven van het leerproces, maar ook leerlingen moet motiveren (bijv. Black & Wiliam, 1998).

Tijd	Informatie over het leerproces (Scriven, 1967)			
	Informatie over het leerproces (Bloom et al., 1971)	die leraren kunnen gebruiken voor onderwijsbeslissingen.		
	Informatie over het leerproces (Sadler, 1989)	die leraren kunnen gebruiken voor onderwijsbeslissingen	en leerlingen kunnen gebruiken om hun leren te bevorderen.	
	Informatie over het leerproces (Black & Wiliam, 1998; Brookhart, 2007, Crooks, 1987)	die leraren kunnen gebruiken voor onderwijsbeslissingen	en leerlingen kunnen gebruiken om hun leren te bevorderen	en ook leerlingen motiveert.
	Groei in definitie			

Figuur 1. Concepties in de definitie van formatief assessment over de tijd (overgenomen uit Sluijsmans et al., 2013 en gebaseerd op Brookhart, 2007, p. 44).

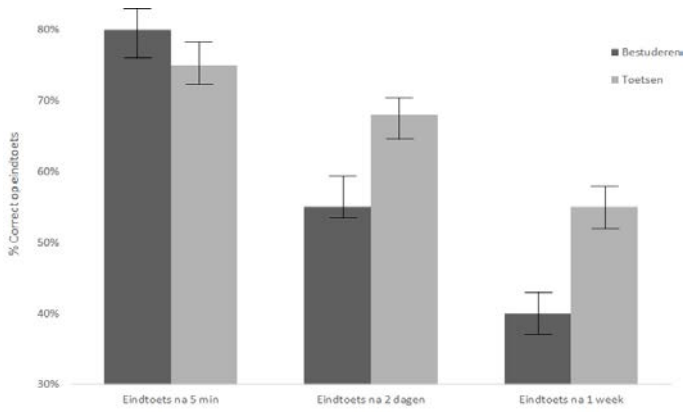
Discussies over het begrip formatief toetsen hebben ook betrekking op de vraag of het bij formatief toetsen gaat om het instrumentarium dat gebruikt wordt (de 'toets') of over het proces (het 'formatieve'), waarin feedback een belangrijke rol speelt. In dat laatste geval wordt formatief toetsen veelal beschreven als 'assessment for learning'. De definitie hiervan is: *"... het proces van zoeken, aggregeren, interpreteren van informatie die studenten en docenten gebruiken om te bepalen waar studenten staan in hun leerproces, waar zij naartoe moeten en op welke manier"* (Assessment Reform Group, 2002). In dit proces zijn verschillende manieren om formatieve informatie te verzamelen (bijvoorbeeld een klassengesprek, exit-ticket, toets etc.) en gaat het er vooral om wat er vervolgens met deze informatie gedaan wordt (Wiliam & Leahy, 2015). Het verzamelen van informatie is dan voornamelijk bedoeld voor het reguleren van het leerproces (Brookhart, 2001). Dochy, Segers, Gijbels, en Struyven (2007) onderscheiden verder pre-, post- en pure effecten van toetsen. Pre-effecten hebben betrekking op de wijze waarop studenten zich voorbereiden op een toets op basis van hun verwachting, Post-effecten hebben betrekking op de wijze waarop bijvoorbeeld feedback bijdraagt aan leren en pure effecten zijn de effecten die het maken van een toets direct heeft op het leren. Het zijn met name deze "pure" effecten waar de literatuur - binnen de geheugenpsychologie waar wij ons op zullen richten - op inspeelt.

In deze reviewstudie gebruiken wij de instrumentele (smalle) definitie: *Een formatieve toets is een toets die leraren en leerlingen informatie geeft over het leerproces en onderdeel is van het leerproces. Het is een digitale of papieren toets met open en/of gesloten vragen die tussentijds gebruikt wordt om het leren te bevorderen.* Een toets die gebruikt wordt om na een leerperiode vast te stellen of kennis en vaardigheden beheerst worden, een zogenoemde eindtoets, kan ook formatief gebruikt worden, maar valt buiten onze definitie van een formatieve toets. Daarnaast vallen andere manieren van informatieverzameling (bijv. een klassengesprek), die wel binnen de brede definitie gebruikt worden, buiten het bereik van deze reviewstudie. Veel gebruikte benamingen voor de formatieve toets die wij hier bedoelen zijn 'oefentoets' of 'diagnostische toets', 'quiz', 'deeltoets' of 'tussentijdse toets'. Voor deze strikte afbakening is gekozen vanwege de wijze waarop de meeste studies naar het testing-effect en spacing effect zijn opgezet. In de onderzoeken wordt namelijk het leereffect van het beantwoorden van toetsvragen op een bepaald moment vergeleken met het herbestuderen

van een boekhoofdstuk. Om de keuze te verduidelijken wordt nu verder ingegaan op de theoretische grondslag uit de geheugenpsychologie, namelijk het *retrieval practice* effect.

1.2. Retrieval practice

Het effect van *retrieval practice* op het leren en onthouden van informatie is in de afgelopen vijftien jaar zeer frequent onderzocht (zie Adesope, Trevisan & Sundararajan, 2017). *Retrieval practice* betekent letterlijk het oefenen met het ophalen van informatie uit het geheugen. Meestal gebeurt dit ophalen van informatie door het beantwoorden van toetsvragen. Er zijn honderden studies die laten zien dat deze vorm van oefening helpt om nieuwe kennis te onthouden (zie Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Adesope et al., 2017 voor overzichten). In eerste instantie werd het effect van retrieval practice vooral onderzocht in de onderwijspraktijk waarbij er weinig controle was qua onderzoeksaanpak (zie Bangert-Drowns, 1999 voor een overzicht). Toen de eerste positieve resultaten van testing en spacing gevonden werden, werd het onderzoek naar deze effecten verplaatst naar het lab waarbij gebruik gemaakt werd van eenvoudige en veelal "onnatuurlijke" leermaterialen (zie Delaney, Verkoeijen, en Spirgel, 2010 voor een overzicht). In een typisch *retrieval practice* experiment bestudeert een deelnemer een aantal woorden of woordparen die hij moet reproduceren op een latere toets. Vervolgens gaat hij ofwel het leermateriaal nog eens bestuderen (de herbestudeer-conditie) ofwel een toets maken over de leerstof, waarbij hij de nieuwe kennis moet ophalen uit zijn geheugen (de *retrieval practice*-conditie). Op een later moment (i.e., 2 dagen of een week) krijgt de deelnemer een eindtoets over het leermateriaal. Dit leidt tot een robuust geheugen-voordeel voor het leermateriaal dat tussentijds is opgehaald uit het geheugen door middel van *retrieval practice* in vergelijking met herbestuderen op een uitgestelde eindtoets (zie Adesope et al., 2017; Rowland, 2014, voor overzichten; zie ook Figuur 2).



Figuur 2. Weergave van het retrieval practice effect (gebaseerd op Roediger & Karpicke, 2006, p. 193).

Dit effect wordt veroorzaakt door een vlakkere vergeetcurve vergeleken met andere strategieën zoals herlezen. Of, met andere woorden, door informatie op de halen uit je geheugen na het leren, vergeet je die informatie minder snel dan wanneer je dezelfde informatie nog eens gelezen zou hebben. Het retrieval practice effect treedt dan ook meestal alleen op wanneer de eindtoets "uitgesteld" is (minimaal 1 dag) en niet direct volgt op de oefentoets (zie Figuur 2).

Een aantal verklaringen - die elkaar niet noodzakelijkerwijs uitsluiten - worden in de literatuur gegeven voor het *retrieval practice* effect (zie Roediger & Karpicke, 2006; Rowland, 2014). Sommige verklaringen richten zich op de moeite die het kost om het leermateriaal op het halen nadat het eerst is bestudeerd. De *theory of disuse* (Bjork & Bjork, 1992) stelt bijvoorbeeld het effect wordt veroorzaakt door de moeite (*effort*) die in de ophaal-poging zit besloten. Dit leidt er bijvoorbeeld toe dat moeilijke toetsen meer bijdragen aan de sterkte waarmee een item is opgeslagen in het geheugen dan gemakkelijke toetsen. Volgens de meer uitgewerkte *elaborative retrieval* hypothese (Carpenter, 2009) zorgt het ophalen van een item uit het geheugen voor activatie van andere items in het geheugen. Deze andere items kunnen vervolgens als aanwijzing (*cue*) gebruikt worden om het doel-item op een later moment op te halen uit het geheugen. De *mediator effectiveness* hypothese richt zich ook op de activatie van *mediators* in het geheugen tijdens *retrieval practice* die op een later moment kunnen helpen bij het ophalen van een item (Pyc & Rawson, 2011). Een meer recente verklaring voor het retrieval practice effect is de *episodic context account*, die stelt dat bij het ophalen van

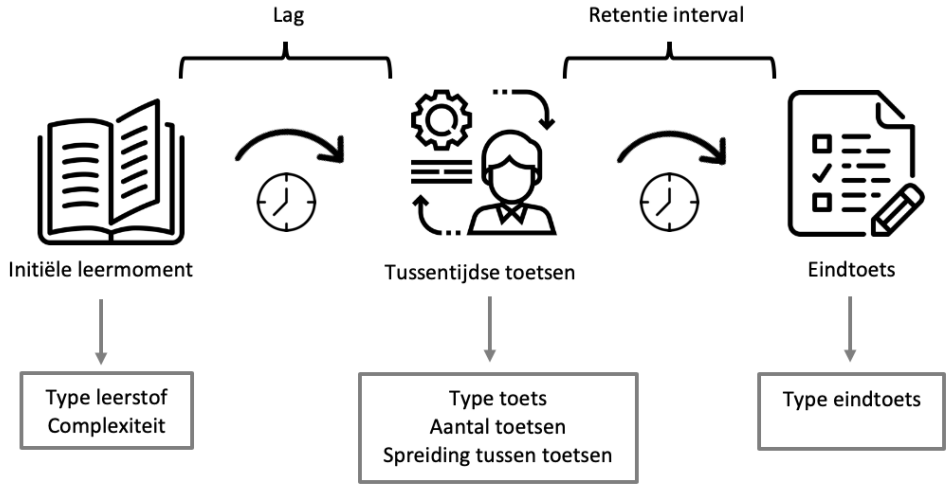
informatie uit het geheugen de contextuele kenmerken die al zijn opgeslagen bij het geheugen-item worden verrijkt met kenmerken van de nieuwe context waarin het item is opgehaald (Whiffen & Roediger, 2017). Dit leidt tot een rijkere representatie in het geheugen, die het gemakkelijker maakt het item op een later tijdstip uit het geheugen op te halen. Een ander soort verklaring voor het effect van *retrieval practice* richt zich op de overlap in het type geheugenprocessen dat plaatsvindt tijdens het oefenen en tijdens het toetsen. Volgens de *transfer-appropriate processing* (TAP) theorie hangt de grootte van het effect van oefenen af van de gelijkheid in het verwerkingsproces tussen de oefensessie en de eindtoets. Bijvoorbeeld, als in de oefensituatie informatie via een herkenningsvraag opgehaald wordt, en in de uiteindelijke toets ook, dan is er meer overlap in de verwerkingsprocessen dan wanneer de eindtoets een open antwoord vereist. Als de overlap groter is, zal het *retrieval practice* effect ook toenemen. Er zijn echter aanwijzingen dat overlap in verwerkingsprocessen niet in alle gevallen leidt tot een groter *retrieval practice* effect (Carpenter & DeLosh, 2005).

Hoewel het *retrieval practice* effect (ook wel aangeduid met het *testing effect*) uit de geheugenpsychologie komt, is het ook uitgebreid onderzocht in de context van het onderwijs (bijv. Dirkx, Kester & Kirschner, 2014; Goossens, Camp, Verkoeijen, Tabbers & Zwaan, 2012; zie Dunlosky et al., 2013 voor een overzicht). Gebleken is dat het eenvoudig te gebruiken is in het onderwijs (Dunlosky et al., 2013), maar dat de effectiviteit van deze oefenstrategie niet goed bekend is bij studenten en docenten (bijv. Bartoszewski & Gurung, 2015; Morehead, Rhodes & Delozier, 2016). Onderzoeken met verschillende typen leermateriaal (bijv. een leerboek, woordenlijst, college of punten op een geografische kaart), met verschillende soorten toetsvragen (kort-antwoord, multiple-choice, matchingvragen, invulvragen), met verschillende populaties (van basisschoolleerlingen tot volwassenen) laten echter zien dat het tussentijds ophalen van informatie leidt tot robuuste effecten op een (uitgestelde) eindtoets (zie bijv. Adesope et al., 2017; Dunlosky et al., 2013; Rowland, 2014). En dat deze effecten te zien zijn op zowel toetsen die puur het onthouden van feiten meten, als ook toetsen die begrip en toepassing meten.

Interessant is dat de manier waarop tussentijdse toetsen voorafgaand aan een eindtoets ingezet worden in *retrieval practice*-experimenten in de onderwijspraktijk grote parallellen vertoont met hoe formatieve toetsen in de onderwijspraktijk worden ingezet. De insteek is echter anders: waar formatieve toetsen meer worden gebruikt als een bron voor

effectieve regulatie van het leerproces, wordt *retrieval practice* ingezet om het onthouden en begrijpen van het leermateriaal te verbeteren (i.e., als leerstrategie). De omvang van het positieve effect van *retrieval practice* op het geheugen hangt vervolgens af van een aantal factoren.

In Figuur 3 wordt een standaard opstelling van een retrieval practice experiment weergegeven en zijn factoren ingetekend die de grootte van het effect op leren bepalen (zie ook Adesope et al., 2017). Een aantal factoren dat de grootte van het *retrieval practice* effect beïnvloedt heeft te maken met de aard van de gebruikte materialen (i.e., leermateriaal en toetsmateriaal). Een belangrijke vraag in de literatuur is bijvoorbeeld welke (onderwijs-relevante) materialen baat hebben bij oefening door middel van *retrieval practice*. Een ander aspect dat een rol speelt, is de vorm van de tussentijdse toets. Hier is in de literatuur de vraag met welk soort tussentijdse toetsen je het grootste leereffect teweeg kunt brengen. Leidt bijvoorbeeld het gebruik van herkenningstaken (zoals bij het herkennen van het juiste antwoord bij *multiple-choice* vragen) tot een *retrieval practice* effect, of is het beter om een *recall* taak (zoals bij een open vragen toets) te gebruiken? In dit kader is ook de aard van de eindtoets belangrijk: op wat voor soort eindtoetsen wordt er een positief effect van *retrieval practice* gevonden? En is het belangrijk dat het type toets dat gebruikt wordt als tussentijdse toets dezelfde vorm heeft als de summatieve eindtoets?



Figuur 3. Visuele weergave van een standaard retrieval practice experiment met factoren die de grootte van het effect beïnvloeden.

Een tweede categorie factoren die de grootte van het *retrieval practice* effect beïnvloedt, heeft te maken met de planning van het leerproces in de tijd. Hoeveel tijd moet er bijvoorbeeld zitten tussen het initiële leermoment (i.e., de fase waarin de lerende de kennis voor het eerst tot zich neemt door bijv. het bestuderen van een boek of bijwonen van een college) en het tussentijds ophalen van het leermateriaal. In de literatuur wordt dit interval aangeduid met de Engelse term *lag*. Een tweede vraag in de literatuur is hoe vaak de tussentijdse toets moet worden afgenomen. Is één keer oefenen voldoende voor een leereffect of moet vaker geoefend worden? Een derde tijdsfactor die de grootte van het *retrieval practice* effect kan beïnvloeden, is de tijd tussen de (laatste) tussentijdse toets en de eindtoets. Dit interval wordt in de literatuur aangeduid met de term *retention interval*. Wat is een ideaal interval voor een goede prestatie op de eindtoets? Voor het beantwoorden van deze vragen is - naast de literatuur over *retrieval practice* - ook de literatuur over een andere geheugenstrategie relevant, namelijk de literatuur over het effect van *distributed practice* (oftewel gespreide oefening). Het *distributed practice* effect wordt in de volgende paragraaf besproken. De kern is dat de wetenschappelijke literatuur over *retrieval practice* en *distributed practice* voor een groot deel antwoord geeft op de bovenstaande vragen, maar de link tussen deze antwoorden en formatief toetsen nog niet eenduidig is gelegd. Het systematisch bestuderen van de antwoorden op deze vragen in het licht van formatief toetsen in de onderwijspraktijk kan concrete richtlijnen opleveren om het leereffect van formatieve toetsen te optimaliseren.

Zoals in de vorige paragraaf is aangegeven ligt de focus van dit onderzoek op de directe effecten van toetsen. Echter, naast het directe effect van het tussentijds ophalen van informatie op leren, worden er (net als in de onderwijskundige literatuur) in de literatuur uit de geheugenpsychologie ook indirecte effecten beschreven van *retrieval practice*. Voorbeelden van dergelijke indirecte effecten zijn meer motivatie, betere inschatting van het eigen leren, betere sturing van het eigen leerproces en afname van de angst voor toetsen (zie Roediger & Karpicke, 2006 voor een overzicht). Deze indirecte effecten komen ook terug in de literatuur rondom formatief toetsen aangezien het daar onder andere gaat om het vergroten van de motivatie van de lerende en het verbeteren van de zelfregulatie. Hier is dus al een heldere overlap zichtbaar tussen de literatuur over *retrieval practice* en de literatuur over formatief

toetsen. Echter, in de literatuur over formatieve toetsing lijkt er nog geen gebruik te worden gemaakt van de kennis over de directe effecten van tussentijds toetsen op leren.

1.3. Distributed practice

Een tweede geheugenstrategie met relevantie voor de inrichting van formatief assessment is *distributed practice*, ofwel gespreid oefenen. Het *distributed practice* effect (ook wel *spacing* effect genoemd) is het fenomeen dat het spreiden van oefening met leermateriaal over de tijd tot beter geheugen leidt voor dat materiaal dan het concentreren van de oefening in één oefensessie. Net als bij *retrieval practice* laat de literatuur over dit fenomeen zien dat het om een robuust effect gaat (zie Cepeda, Pashler, Vul, Wixted & Rohrer, 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Dunlosky et al., 2013; Kang, 2016). Ook dit fenomeen vindt zijn oorsprong in laboratorium studies. In een typisch experiment worden woorden (bijv. woord A, B en C) in een leersessie op verschillende manieren herhaald: in de *massed practice* conditie wordt de herhaling van elk woord achter elkaar gepland in de tijd (bijv. AAABBBCCC-eindtoets). In de *distributed practice* conditie wordt de herhaling van woorden gespreid in de tijd (bijv. ABCABCABC-eindtoets). Hoewel de tijd die besteed wordt aan het oefenen gelijk is in beide condities, en ook het aantal herhalingen van elk woord niet verschilt tussen de condities, is het geheugen voor woorden in de *distributed practice* conditie beter dan in de *massed* conditie.

Twee verklaringen voor dit fenomeen zijn dominant in de literatuur over *distributed practice*. De eerste richt zich op het gegeven dat de overlap tussen de context waarin het materiaal is geleerd en de context waarin het materiaal moet worden opgehaald uit het geheugen de kans bepaalt dat een item succesvol wordt herinnerd. Volgens de *encoding variability* theorie worden items die gespreid geoefend worden als gevolg van de spreiding in verschillende contexten geoefend, terwijl items die achter elkaar geoefend worden maar in één beperkte context worden geoefend (bijv. Glenberg, 1979). Kenmerken van die verschillende contexten worden bij het item opgeslagen in het geheugen en kunnen later als aanwijzing (*cue*) dienen om het item te herinneren. De rijkere context zorgt daarmee voor meer *cues* in het geheugen die helpen om de items die gespreid zijn geoefend later uit het geheugen op te halen in de eindtoets. De tweede theorie breidt dit idee verder uit. Volgens de *study-phase retrieval* theorie (bijv. Hintzman, Summers & Block, 1975) worden de rijkere contextuele

elementen pas waardevol voor het geheugen als de lerende (automatisch) de eerdere blootstelling aan het item ophaalt tijdens de oefening met het item. Alleen als de vorige blootstelling aan het item wordt opgehaald uit het geheugen worden de nieuwe context-elementen aan de geheugen-representatie toegevoegd en is de kans groter dat het item op een later moment opgehaald kan worden uit het geheugen.

Net als retrieval practice, werkt *distributed practice* niet alleen in het lab, maar ook in de onderwijspraktijk en is het effect in een grote variatie aan onderwijs-contexten onderzocht en eenvoudig in te zetten in de onderwijspraktijk (zie Dunlosky et al., 2013; Gerbier & Toppino, 2015, voor een overzicht). Een studie bij leerlingen op de basisschool liet bijvoorbeeld zien dat het oefenen met nieuwe woordenschat gespreid over drie dagen tot beter geheugen leidde voor de nieuwe woorden op een uitgestelde toets dan het concentreren van de oefening op één dag (Goossens, et al., 2012). Kinderen leerden bijvoorbeeld op maandag 15 nieuwe woorden en oefenden vervolgens drie keer met de nieuwe woorden ofwel door ze drie keer te oefenen op één dag (de *massed* conditie) ofwel door ze drie keer te oefenen op drie verschillende dagen (de *spaced* conditie, bijv. op dinsdag, woensdag en donderdag). Op een eindtoets na 1 week en na 5 weken bleek het geheugen voor de woorden die gespreid waren geoefend beter te zijn.

Het effect van *distributed practice* kent meerdere dimensies (zie Figuur 3). Ten eerste kan de tijd tussen het initiële leermoment en de (eerste) oefening worden gevarieerd (de *lag*). De relevante vraag hier is dan welke *lag* zorgt voor een optimaal leereffect. Daarnaast kan de tijd tussen de verschillende oefensessies worden gevarieerd als het er meer dan één is (ook wel: *spacing*). Op het gebied van de *spacing* van oefensessies is bijvoorbeeld onderzocht of het een voordeel heeft om de tijd tussen *retrieval practice* sessies langzaam te vergroten (*expanding spacing*) of om de tijd tussen oefensessie gelijk te houden (*equal-interval spacing*) (bijv. Kang, Lindsey, Mozer & Pashler, 2014; Karpicke & Roediger, 2007). Ook kan de tijd tussen de laatste oefensessie en de eindtoets worden gevarieerd (het *retention interval*). Deze laatste variatie geeft ons inzicht in de relatie tussen het schema van gespreide oefening en het geheugen voor het leermateriaal op verschillende momenten later in de tijd. De relevante vraag hier is wat de optimale plaatsing in de tijd is van de oefensessies voorafgaand aan de eindtoets.

Door het systematisch onderzoeken van de kennis over de effecten van deze factoren binnen de *retrieval practice* en *distributed practice* literatuur, kunnen concrete richtlijnen

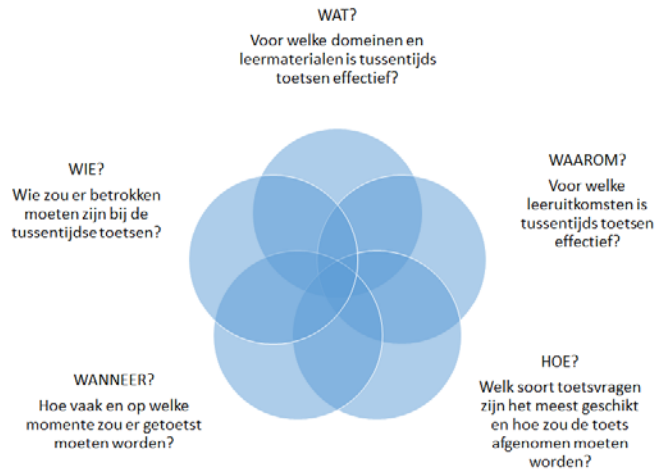
worden afgeleid voor de vorm, plaatsing in de tijd etc. van formatieve toetsen. Voor het systematisch onderzoeken van de ontwerprichtlijnen, is gebruik gemaakt van het ontwerpproces van toetsen zoals beschreven door Stiggins en Conklin (1952) en Van der Vleuten en Driessen (2000).

1.4. Ontwerprichtlijnen voor formatief toetsen

Het ontwerpproces voor (formatieve) toetsen bestaat onder andere uit het bepalen van het doel van de toets, het in kaart brengen van datgene wat men wilt toetsen door het opstellen van leerdoelen en een toetsmatrijs en het bepalen van een geschikte toetsvorm op basis van de geformuleerde leerdoelen. Stiggins en Conklin (1952) en Van der Vleuten en Driessen (2000) hebben dit ontwerpproces beschreven aan de hand van vijf vragen die een leraar zich zou moeten stellen in de ontwerpfase:

1. Wat beoordelen? Voor welk soort leermateriaal wil je de toets gebruiken?
2. Waarom beoordelen? Wat wil je met de toets meten, wat is het doel waarvoor je toetst?
3. Hoe beoordelen? Welke vraagtypen en hulpmiddelen worden ingezet om informatie te verzamelen?
4. Wanneer beoordelen? Tussentijds, continu?
5. Wie betrekken bij het beoordelen? docenten, studenten?

Deze vijf vragen zijn te koppelen aan de factoren die eerder beschreven zijn als bepalend voor (de grootte van) het *retrieval practice* en *distributed practice* effect (zie Figuur 4) en vormen daarmee het analysekader in deze literatuurstudie voor het formuleren van ontwerprichtlijnen voor formatieve toetsen vanuit de geheugenpsychologie.



Figuur 4. Analysekader gebaseerd op Van der Vleuten en Driessen (2000)

1.5. Onderzoeksvraag

Zoals boven beschreven is er veel bekend over factoren die invloed hebben op de grootte van het *retrieval practice* en *distributed practice effect*. Aangezien er veel overeenkomsten zijn tussen retrieval practice, distributed practice en formatief toetsen kunnen de inzichten uit dit type onderzoek dus zeer informatief zijn voor het ontwerpen van formatieve toetsen. Echter, (naar beste weten) is er nog geen onderzoek gedaan waarin tussentijdse toetsen gebruikt zijn die ontworpen zijn op basis van de geheugenpsychologie. Ook bieden de onderzoeken vanuit de geheugenpsychologie die tot dusver zijn uitgevoerd geen totaaloverzicht van ontwerprichtlijnen voor tussentijdse toetsen. De centrale onderzoeksvraag voor deze review is daarom dan ook: *Op welke manier kunnen principes van retrieval practice en distributed practice gebruikt worden voor het effectief ontwerpen en inzetten van formatieve toetsen in de onderwijspraktijk?*

Het doel is te komen tot concrete ontwerprichtlijnen voor effectief gebruik van formatieve toetsen in de praktijk gebruik makend van studies uitgevoerd in de geheugenpsychologie. De volgende deelvragen zijn daarbij geformuleerd op basis van het kader in Figuur 4:

1. *Voor welke domeinen en leermaterialen is tussentijds toetsen effectief? (wat?)*
2. *Voor welke leeruitkomsten is tussentijds toetsen effectief? (waarom?)*
3. *Welk soort toetsvragen zijn het meest geschikt en hoe zou de toets gegeven moeten worden? (hoe?)*
4. *Hoe vaak en op welke momenten zou er getoetst moeten worden? (wanneer?)*
5. *Wie zou betrokken moeten zijn bij de toetsafname? (wie?)*

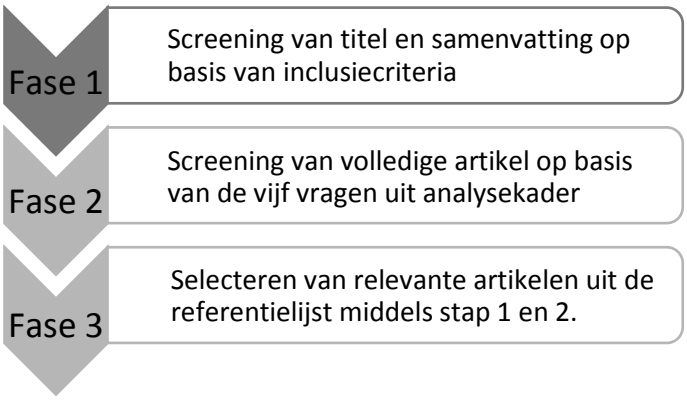
Als basis voor het beantwoorden van deze vraag is gebruik gemaakt van literatuur uit twee onderzoeksstromingen: de literatuur uit de geheugenpsychologie over *retrieval practice* en *distributed practice*. Ook is er gekeken naar de onderwijskundige literatuur over formatief toetsen om te zien of daar niet al verwezen wordt naar de geheugenpsychologie.

2. Methode

2.1. Literatuursearch en selectiecriteria

Om de onderzoeksvraag te beantwoorden is een systematische literatuurstudie uitgevoerd in twee databases, te weten: Web of Science en Ebsco Host. Engelse en Nederlandse zoektermen werden ingedeeld in drie categorieën: *retrieval practice*, formatief toetsen en richtlijnen. Er werd verder gefilterd op Engels of Nederlandstalige empirische peer-reviewed artikelen, overzichtsartikelen, rapportages, en proefschriften gepubliceerd tussen 2009 en 2018[1].

Initieel leverde het zoeken 405 resultaten verdeeld over de tw110ee databases op. Na het verwijderen van dubbelingen, bleven er 275 artikelen over. Via een drie-traps filtering (zie Liberati al., 2009) werd dit aantal teruggebracht tot 71 en werden 40 artikelen aan de hand van verwijzingen en de referentielijst toegevoegd (sneeuwbalmethode). De fases en gebruikte inclusiecriteria zijn beschreven in Figuur 6.



Figuur 6. Selectieproces

2.2. Codeerschema

Alle 110 artikelen in de reviewstudie werden gecodeerd met behulp van een schema bestaande uit drie onderdelen: beschrijvende informatie, antwoord op deelvragen en kwaliteit. In het eerste onderdeel werd de volgende informatie opgenomen: referentie naar het artikel,

[1] Deze afbakening is gekozen vanwege het groot aantal review studies van de afgelopen jaren die de eerdere studies al meenemen.

samenvatting, deelnemers (leeftijd, onderwijsniveau, steekproefgrootte), resultaten, conclusie, generaliseerbaarheid, implicaties en limitaties zoals beschreven door de auteur. In het tweede onderdeel werd informatie opgenomen uit het artikel dat antwoord gaf op een van de vijf ontwerpvragen (zie onderzoeksvragen). Deze vragen zijn gebaseerd op het raamwerk gepresenteerd in Figuur 4. Daarnaast werden kwaliteitsaspecten van de studies beschreven (zie Tabel 1).

Tabel 1.

Beschrijving van het analysekader

Beschrijvende informatie	Referentie Samenvatting Deelnemers (leeftijd, onderwijsniveau, aantal) Resultaten Conclusie Generaliseerbaarheid Implicaties Limitaties
Antwoord onderzoeksvraag	Voor welke domeinen en leermaterialen is tussentijds toetsen effectief? Hoe vaak en op welke momenten zou er getoetst moeten worden? Voor welke leeruitkomsten is tussentijds toetsen effectief? Wie zou betrokken moeten zijn bij de toetsafname en in welke vorm? Welk soort toetsvragen zijn het meest geschikt en hoe zou de toets gegeven moeten worden?
Kwaliteit van het onderzoek	Adequaatheid van de onderzoeksmethode Selectie van onderzoekseenheden Wijze van dataverzameling Wijze van data-analyse Beperkingen van het onderzoek

2.3. Data-analyse

De data-analyse is uitgevoerd in vier stappen. In stap 1 werd voor elke ontwerpvraag de relevante informatie uit de verschillende artikelen uit de overzichtstabel gehaald en bij elkaar gezet in een document. In stap 2 werden de samengevoegde teksten doorgenomen op expliciet gegeven richtlijnen (i.e., richtlijnen die direct voortvloeien uit het beantwoorden van de onderzoeksvraag) en impliciete richtlijnen (i.e., richtlijnen die niet direct voortvloeien uit het beantwoorden van de onderzoeksvraag maar wel door de auteur beschreven worden). In stap 3 werden de resultaten doorgenomen door een tweede onderzoeker, om na te gaan of er geen informatie over het hoofd gezien is en om na te gaan of de meer impliciete richtlijnen wel daadwerkelijk af te leiden zijn uit de onderzoeken. Hierbij werd niet alleen de overzichtstabel geraadpleegd, maar werd ook het volledige artikel weer geraadpleegd. In stap 4 is gekeken of er uit de informatie, per deelvraag, richtlijnen gedestilleerd kunnen worden door de informatie in een aantal iteraties samen te vatten. Deze richtlijnen zijn vervolgens besproken met alle auteurs en werden daarna aangescherpt door de eerste auteur.

3. Resultaten

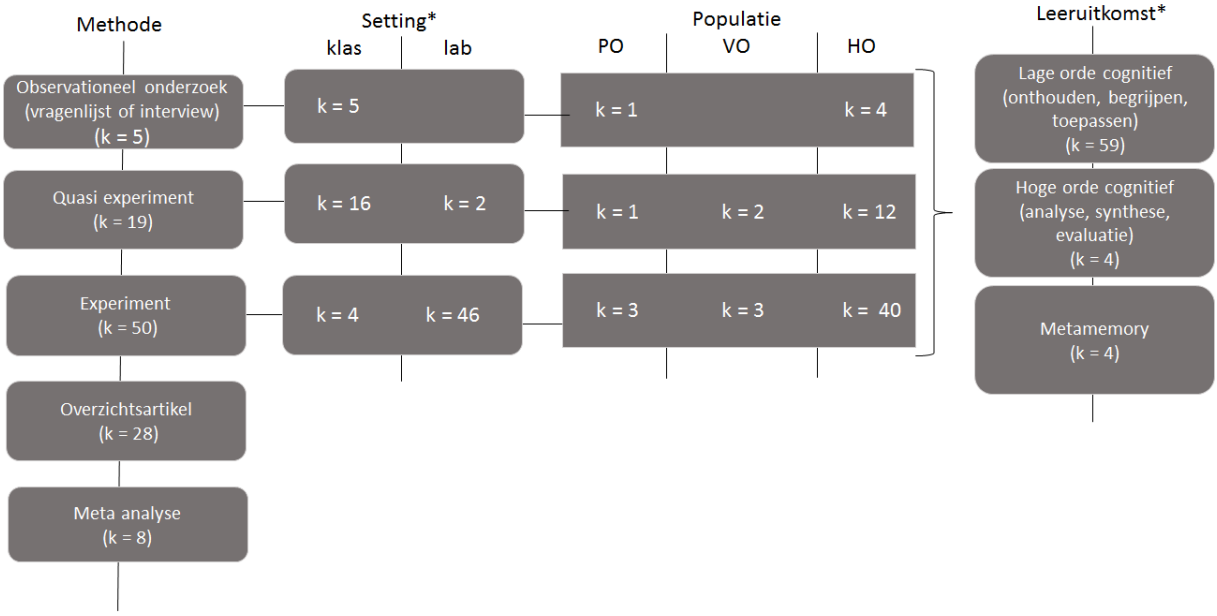
Het doel van deze overzichtsstudie was te komen tot concrete ontwerprichtlijnen voor formatieve toetsen op basis van de literatuur over *retrieval practice* en *distributed practice*. Voordat we de richtlijnen zullen presenteren, zullen we aangeven wat de reikwijdte van de richtlijnen is en op welke wijze ze gelezen moeten worden. Daarvoor geven we een beknopt overzicht van de gevonden studies.

3.1. Algemeen

Van alle artikelen die gevonden zijn op basis van de gebruikte zoektermen, vertrekken 83 artikelen vanuit de *retrieval practice* of *distributed practice* literatuur. De andere artikelen vertrekken onder andere vanuit de literatuur over formatief toetsen (bijv. Bennett, 2011; Kwan, 2011; Sluijsmans, et al., 2013), quizzing met computers (bijv. Mayer et al., 2009) of *classroom testing* (bijv. Bangert-Drowns, Kulik, & Kulik, 1991).

Verreweg het meeste onderzoek uit de *retrieval practice* en *distributed practice* literatuur is uitgevoerd met een experimenteel design in het lab bij universitaire studenten (veelal uit een proefpersonen-bestand). Ook richten de meeste onderzoeken zich op lage orde cognitieve vaardigheden (onthouden, begrijpen en toepassen in vrijwel identieke opgaven; zie Figuur 7). Ook wordt er vaak relatief eenvoudig materiaal gebruikt (woordparen, tekstpassages) en zijn de retentie intervallen kort (meestal 1 week). De omstandigheden waarin dit onderzoek is uitgevoerd zijn dus anders dan de situatie in de onderwijspraktijk, waar vaak lange retentie intervallen nagestreefd worden (het doel is immers dat de aangebrachte kennis beklijft), en in de eindtoets vooral verwacht wordt dat leerlingen of studenten ook in staat zijn andere informatie dan in de tussentijdse toets, op te halen en te gebruiken in complexe taken (analyse, synthese etc). Gelukkig zijn er ook een aantal studies (veelal quasi experimenteel) waarbij wel onderwijs relevant materiaal gebruikt werd zoals de inhoud van colleges. (Foss & Pirozollo, 2017). Of studies waarin het retentie interval langer was dan 7 dagen. In een studie van Cepeda, et al., 2008 werd zelf en retentie interval van 350 gebruikt. Ook zijn er studies waarin leerlingen uit het middelbaar onderwijs (Küpper-Tetzel, Erdfelder, & Dickhauser, 2014; McDaniel, Thomas,

Agarwal, McDermott, & Roediger, 2013) of het basisonderwijs meededen (Fletcher & Shaw, 2012; Vlach & Sandhofer, 2012).



Figuur 7. Overzicht van het type artikelen dat gevonden is (*k*= het aantal artikelen).

*niet alle artikelen rapporteren het design of de uitkomstmaten duidelijk genoeg.

Verder is het belangrijk bij het lezen van de richtlijnen in het achterhoofd te houden dat deze richtlijnen voornamelijk voortkomen uit onderzoek dat is uitgevoerd vanuit de geheugenpsychologie waarbij het primaire doel van de tussentijdse toetsen het vergroten van het *retrieval success* is (i.e., zoveel mogelijk antwoorden goed op de toets) voor de opgehaalde informatie. De *retrieval practice* en *distributed practice* literatuur richten zich veel minder op indirecte effecten van tussentijds toetsen zoals het stimuleren van studiegedrag, het verhogen van motivatie en zelfregulatie (zie voor een van de uitzonderingen Ariel & Karpicke, 2017). Bij het maken van ontwerpkeuzes voor het inzetten van tussentijdse toetsen in de reële onderwijspraktijk is het belangrijk om het doel van de toets goed voor ogen te houden. Roediger, Putnam, en Smith (2011) geven overigens een overzicht van zowel de directe als indirecte effecten van *retrieval practice*, maar niet wat dit betekent voor ontwerp van de formatieve toets.

3.2. De richtlijnen

Hieronder presenteren we negen richtlijnen voor formatieve toetsen die voortkomen uit de reviewstudie over *retrieval practice* en *distributed practice* uit de geheugenpsychologie. De inhoud en volgorde van de richtlijnen volgen de ontwerpvragen zoals in de onderzoeksvragen beschreven.

1. Gebruik formatieve toetsen in verschillende domeinen en bij verschillende soorten leermaterialen (bijv. teksten, woordjes, sommen, presentaties en video's) om leren te stimuleren;

   Onderzoek laat zien dat retrieval practice en distributed practice effectief zijn voor een breed scala aan leermateriaal (zie bijv. Adesope et al., 2017; Pan & Rickard, 2018) zoals teksten, woordparen en hoorcolleges in allerlei domeinen (bijv. wiskunde, geschiedenis, vreemde talen, biologie, geneeskunde, en psychologie). Het onderzoek laat zien dat effect het grootst is voor materiaal waarin de verschillende woorden of concepten nauw met elkaar verweven zijn (bijv. een tekst over één onderwerp; zie bijv. Adesope et al., 2017; Karpicke, 2017) en bij wiskunde (Adesope et al., 2017). Vervolgonderzoek naar het effect op complexe leermaterialen (bijvoorbeeld uitgewerkte voorbeelden) is nog nodig (zie Van Gog & Sweller, 2015; Karpicke & Aue, 2015).

2. Gebruik formatieve toetsen in elk geval voor onthouden, begrijpen, en toepassen van informatie;

   Het doel van retrieval en distributed practice is door het tussentijds ophalen van informatie uit het geheugen de kans te vergroten dat die informatie op een later moment herinnerd kan worden. Recent onderzoek heeft ook gekeken of dit effect heeft op het beantwoorden van begripsvragen en toepassingsvragen waarin dezelfde soort informatie (woordjes, concepten, definities) in een ietwat andere vraagstelling getoetst wordt (transfer). De effecten van tussentijds toetsen zijn ook bij dit soort leeruitkomsten positief (zie Adesope et al., 2017 en Pan & Rickard, 2018 voor een overzicht). De laatste jaren zijn er ook enkele studies (bijv. Agarwal, 2018; Johnson & Mrowka, 2010; Narloch,

Garbin, & Turnage, 2006) uitgevoerd die gekeken hebben naar vragen op het niveau van analyse, evaluatie en synthese (Bloom et al., 1971). Voor dit niveau van vragen lijkt er ook een positief effect te zijn, maar aanvullend onderzoek is nodig om de robuustheid daarvan vast te stellen.

3. Stem het niveau en de inhoud van de formatieve toets af op de eindtoets (*constructive alignement*).

Een belangrijke voorwaarde is dat er een goede overeenstemming is tussen de tussentijdse toets en de eindtoets (Fiorella & Mayer, 2016; Kou & Simon, 2009; Stenlund, Sundström, & Jonsson, 2016) in onder andere de inhoud (welke informatie uit het leermateriaal wordt er getoetst?) maar vooral ook het niveau (onthouden, begrijpen etc; Agarwal, 2018). De tussentijdse toets moet de leerlingen dus goed voorbereiden op wat zij moeten kennen (inhoud) en kunnen (herkennen, beschrijven, toepassen) op de eindtoets (bijv. Gibbs & Simpson, 2005). De studie van onder andere Agarwal (2018) laat bijvoorbeeld zien dat wanneer de eindtoets gericht is op hogere orde cognitieve vaardigheden (analyse, synthese, evaluatie), de vragen in de tussentijdse toets ook op dit niveau gesteld moeten worden om effectief te kunnen zijn.

4. Kies voor een combinatie van open en gesloten vragen (zgn. hybride toetsen) bij formatieve toetsen.

Een aantal studies laat zien dat het gebruik van kort-antwoord-vragen, waarin de deelnemer actief informatie moet ophalen uit het geheugen beter is dan multiple-choice-vragen waarin alleen het herkennen van het goede antwoord voldoende is (bijv. Kang, McDermott, & Roediger, 2007). Er zijn echter ook studies die precies het tegenovergestelde laten zien (zie Roediger & Karpicke, 2006; Karpicke, 2017). De tegengestelde resultaten hangen grotendeels samen met het al dan niet geven van feedback en hoe goed de leerlingen de tussentijdse toetsen maken. Wanneer tussentijds correctieve feedback gegeven wordt zijn kort-antwoord-vragen vaak het beste (Kang, et al., 2007; Smith & Karpicke, 2014). Dit is deels te verklaren door het feit

dat open vragen moeilijker zijn en dus vaak minder goed worden beantwoord. Maar, open vragen zorgen wel voor een 'actievere' vorm van retrieval practice vergeleken met multiple-choice-vragen, die meer op herkenning dan op retrieval zijn gebaseerd. Door feedback toe te voegen wordt er gecompenseerd voor de lage initiële retrieval maar blijven de voordelen van de meer actieve retrieval wel bestaan. Anderzijds, scoren leerlingen en studenten beter op een tussentijdse multiple-choice toets waarbij ze het goede antwoord kunnen herkennen (lagere orde vaardigheden). Het voordeel van multiple-choice toetsen is dan dat er meer informatie in dezelfde tijds getoetst kan worden. Om de voordelen van beide toetsvormen te combineren, wordt er in recent onderzoek gebruik gemaakt van zogenaamde hybride toetsen die bestaan uit een combinatie van open en gesloten vragen met feedback. Dit soort toetsen laten tot nu toe de grootste effecten zien (zie voor een overzicht Adesope et al., 2017; Jensen et al., 2013; Pan & Agarwal, 2018).

5. Als je formatief toetst, zorg dan dat je in de feedback het goede antwoord geeft;

Er is veel onderzoek gedaan naar het effect van feedback na oefentoetsen binnen de retrieval practice en distributed practice literatuur. Feedback bestaat daar meestal (enkel) uit het geven van het goede antwoord na een toets(vraag). Er wordt niet toegelicht waarom het antwoord goed of fout is. Overzichtsstudies (bijv. Roediger & Butler, 2011; Rowland, 2014) laten zien dat het geven van feedback grotere effecten van tussentijds toetsen oplevert, dan wanneer er geen feedback gegeven wordt. Dit is met name het geval wanneer de tussentijdse toets slecht gemaakt wordt (i.e., er nog niet genoeg geleerd is voordat de toets gemaakt werd). Dan is het geven van feedback het meest belangrijk. Verder is het bij multiple-choice vragen altijd van belang feedback te geven zodat eventuele misconcepties (i.e., het foute antwoord) gecorrigeerd worden (zie Roediger & Butler, 2011). Bij multiple-choice vragen kan namelijk het foute antwoord "herkend" worden maar zal de student dit niet duidelijk als "foute" antwoord herkennen waardoor de kans dat hij dit eerder gekozen antwoord nogmaals kiest, groot is. In de eindtoets maakt de student dan dezelfde soort fouten.

Naar de timing van het type feedback is ook veel onderzoek gedaan. Uit het onderzoek komt naar voren dat zowel directe feedback als uitgestelde feedback helpt (Butler &

Roediger, 2008), maar dat uitgestelde feedback (bijv. aan het einde van de toets in plaats van na elke vraag) grotere effecten laat zien (zie bijv. Roediger & Butler, 2011 en Rowland, 2014 voor overzichten).

6. Zet een formatieve toets pas in na een initiële leerfase.

Onderzoek vanuit de geheugenpsychologie laat zien dat retrieval practice effecten minder groot zijn wanneer het leermateriaal nog heel moeilijk is voor de leerlingen of studenten, dus als ze weinig vragen goed kunnen beantwoorden (Pyc & Rawson, 2009; Rowland, 2014). De uitvoerige meta-analyse van Rowland (2014) toont bijvoorbeeld aan dat wanneer 50% of minder van de antwoorden op de formatieve toets goed beantwoord zijn, er geen retrieval practice effect optreedt. Het is dus - vanuit geheugen-psychologisch perspectief gezien - belangrijk toetsen vooral te gebruiken na een succesvolle initiële leerfase om er zeker van te zijn dat er genoeg informatie kan worden opgehaald uit het geheugen bij de tussentijdse toets. Vanuit de literatuur over formatief toetsen wordt echter juist aanbevolen om een formatieve toets juist voorafgaand aan een leerfase te gebruiken, om voorkennis zichtbaar te maken of te activeren en zo de leerinhoud beter af te stemmen op het kennisniveau van de leerling (Bell & Cowie, 2001; Black & Wiliam, 2009). Hier dient de toets duidelijk een ander doel dan in de geheugenpsychologie. Wanneer het erom gaat het leereffect van de formatieve toets te vergroten, is de richtlijn de toets pas in te zetten als er voldoende informatie opgehaald kan worden uit het geheugen. Als het doel is het leerproces gerichter te sturen of af te stemmen op de lerende, kan de toets ook vooraf ingezet worden.

7. Toets dezelfde stof minstens één keer maar maximaal drie keer.

Twee review studies laten zien dat één keer tussentijds toetsen een even groot of zelfs groter effect heeft dan meerdere keren tussentijds toetsen (Adesope et al., 2017; Rowland, 2014). Daartegenover staat dat in een aantal individuele studies wordt gevonden dat het gebruik van meerdere toetsen het vergeten sterker vertraagt dan één

enkele toets (Karpicke & Roediger, 2008; Pyc & Rawson, 2009; Roediger & Karpicke, 2006). Daarom adviseren sommige onderzoekers juist om driemaal tussentijds te toetsen (Ariel & Karpicke, 2018). De beperkte of afwezige bijdrage van additionele tussentijdse toetsen aan het retrieval-practice effect in review studies wordt mogelijk gemaskeerd doordat in de herstudie-conditie (waarmee wordt vergeleken) alle items worden versterkt, terwijl in de retrieval-practice conditie alleen items worden versterkt die succesvol worden opgehaald uit het geheugen. Het geven van feedback na retrieval practice kan dit probleem verhelpen, maar niet in alle studies wordt deze feedback gegeven. Kortom, de grootste bijdrage voor leren zit in de eerste herhaling en herhaald toetsen is de moeite waard, maar de grootte van de bijdrage van elke additionele toets wordt steeds minder.

8. Spreid de formatieve toetsen uit over de tijd.

Een zeer robuuste bevinding in de literatuur naar het distributed practice effect is dat het effectiever is voor het geheugen van leermateriaal op de lange termijn om de oefeningen met leermateriaal te spreiden in de tijd en ze niet achter elkaar te plannen (zie bijv. Cepeda et al., 2006; Cepeda et al., 2008; Dunlosky et al., 2013). Vervolgens, als er meerdere tussentijdse toetsen worden gebruikt, is er nog de keuze hoe deze tussentijdse toetsen ten opzichte van elkaar in de tijd worden gepland. In een expanding spacing schema neemt de tijd tussen de tussentijdse toetsen toe, terwijl in een equal-interval spacing schema de tijd tussen de tussentijdse toetsen gelijk blijft (bijv. Balota, Duchek, & Logan, 2007). In de literatuur wordt er geen eenduidig voordeel gevonden voor één van deze twee schema's op een eindtoets (zie bijv. Karpicke & Roediger, 2007; Kang, et al., 2014). Wel wordt gevonden dat de prestaties op de tussentijdse toetsen beter is in een expanding spacing schema dan in een equal-interval spacing schema (Kang et al., 2014; Karpicke, 2004), wat wellicht motiverend kan werken voor de lerende (Balota et al., 2007). Een aantal onderzoekers suggereert dat het vooral belangrijk is dat de eerste tussentijdse toets uitdagend moet zijn (i.e., moeite moet kosten, zonder dat het ten koste gaat van de prestatie) en daarom niet te dicht op het initiële leermoment gepland moet worden (Karpicke & Roediger, 2007; Kang et al.,

2014). Hoe de daarop volgende tussentijdse toetsen in de tijd worden gepland is dan van minder belang.

9. Begin niet vlak voor de summatieve toets met het maken van formatieve toetsen, maar gebruik de 20% regel.

Er is geen vaste, ideale tijd tussen het initiële leermoment en de tussentijdse toets (de lag). Er lijkt echter wel een relatie te zijn tussen lag en de tijd tussen de tussentijdse toets en de eindtoets (het retentie-interval, Cepeda et al., 2006; Cepeda et al., 2008; Kang, 2016). Grofweg en enigszins tentatief kan gesteld worden dat de lag 10-20% moet zijn van het retentie-interval. Als bijvoorbeeld 10 dagen na het initiële leermoment een summatieve toets wordt afgenomen, dan zou het advies zijn om 1-2 dagen na het initiële leermoment een tussentijdse toets af te nemen. Het is wel de vraag of bij grote intervallen (bijv. maanden) tussen initieel leermoment en eindtoets dit ook opgaat, omdat dan de lag zo groot is dat de lerende moeite zal hebben zich het leermateriaal te herinneren. De 20% regel is daarom een goed uitgangspunt, maar het leereffect van de eerste tussentijdse toets hangt niet alleen af van de lag en wordt mede bepaald door bijvoorbeeld het succes van de retrieval poging, hoe goed het leermateriaal initieel is geleerd, de complexiteit van het leermateriaal en het type toets dat wordt gebruikt. Een andere relevante bevinding is dat het effect van retrieval practice groter is bij retentie-intervallen die groter zijn dan 1 dag (Adesope et al., 2017; Rowland, 2014). Ook in de literatuur over het distributed practice effect wordt dit teruggevonden: items waarvan de oefening gespreid is in de tijd worden beter onthouden na een langer retentie-interval (zie Balota et al., 2007). Vandaar dat het beter is om niet vlak voor een summatieve toets te beginnen met maken van tussentijdse toetsen.

4. Conclusies en discussie

Formatieve toetsen worden in het onderwijs veel gebruikt. In deze overzichtsstudie stond de vraag centraal op welke manier het onderzoek naar *retrieval practice* en *distributed practice* bij kan dragen aan het effectief ontwerpen en inzetten van formatieve toetsen in de

onderwijspraktijk. Het doel was een met onderzoek gefundeerde basis te leggen voor ontwerprichtlijnen voor formatief toetsen. Om deze vraag te beantwoorden is een systematische literatuurstudie uitgevoerd resulterend in 110 artikelen van waaruit we negen ontwerprichtlijnen hebben kunnen formuleren:

1. Gebruik formatieve toetsen in verschillende domeinen en bij verschillende soorten leermaterialen (bijv. teksten, woordjes, sommen, presentaties en video's) om leren te stimuleren;
2. Gebruik formatieve toetsen in elk geval voor onthouden, begrijpen, en toepassen van informatie;
3. Stem het niveau en de inhoud van de formatieve toets af op de eindtoets;
4. Kies voor een combinatie van open- en gesloten vragen bij formatieve toetsen;
5. Als je formatief toetst, zorg dan dat je in de feedback het goede antwoord geeft;
6. Zet een formatieve toets pas in na een initiële leerfase;
7. Toets dezelfde stof minstens één keer maar maximaal drie keer;
8. Spreid de toetsen uit over de tijd;
9. Begin niet vlak voor de summatieve toets met het maken van formatieve toetsen maar gebruik de 20% regel;

Zoals hierboven aangegeven zijn deze richtlijnen ingegeven door onderzoek vanuit de geheugenpsychologie. De insteek bij dit type onderzoek is anders en de focus ligt (veel meer) op de directe effecten van formatief toetsen. Dit leidt op sommige punten wel voor onbeantwoorde vragen. Een van de voorbeelden is dat de focus van veel onderzoeken tot dusver gericht was op de lagere cognitieve vaardigheden zoals onthouden, begrijpen en toepassen. Het onderzoek naar voor het onderwijs ook zeer relevante leeruitkomsten zoals analyse, synthese en evaluatie staat nog in de kinderschoenen. Echter, de resultaten tot dusver zijn wel zeer veelbelovend. Het is dus interessant om het onderzoek in deze richting verder te vervolgen. Daarnaast is juist het *distributed practice* effect ook bij vaardigheden veelvuldig gebruikt (bijvoorbeeld door in verschillende tijdsintervallen te oefenen; e.g., Verdaasdonk, Stassen, van Wijk & Dankelman, 2007). Een ander dilemma komt voort uit het feit dat er in de onderzoeken vaak sprake is van één 'instructie of leermoment', terwijl in het onderwijs

gedurende een langere periode voortdurend nieuwe informatie aangeboden wordt; in een collegereeks, met zelfstudie etc. Het is dus niet zo dat leerlingen en studenten vanaf één moment oefentoetsen gaan maken over de 'te beheersen leerstof', maar er eigenlijk elke dag nieuwe stof bij komt. De vraag is hoe hier in het onderwijs het beste mee omgegaan kan worden. Tot op heden zijn er nog maar weinig studies gedaan die op deze wijze zijn opgezet en naar een heel semester of schoolkwartaal kijken. Een uitzondering is de studie van Foss en Pirozello (2017). In hun onderzoek is gekeken naar frequent toetsen waarbij dezelfde leerstof terug kwam in een toets waar ook nieuwe leerstof in getoetst werd (i.e., cumulatieve toetsing) tijdens een heel semester. Daarbij is gevarieerd in de toetsvorm en de 'stakes' van de toetsen. De resultaten laten zien dat frequent toetsen leidt tot betere toetsresultaten op de eindtoets. Ook Kerdijk, Cohen-Schotanus, Mulder, Muntinghe, & Tio (2015) gebruiken cumulatief toetsen om tijdens een heel semester toetsen in het curriculum te integreren. Ook hij vindt positieve effecten van cumulatief toetsen en verklaart dit onder andere door het feit dat studenten veel regelmatiger studeren tijdens het semester en niet al het leerwerk uitstellen tot de laatste week. Een conclusie van zijn proefschrift is dat cumulatief toetsen ervoor zorgt dat initiële laagscoorders hun scores verbeteren door cumulatief toetsen zonder dat dit de scores van initiële hoogscoorders negatief beïnvloedt. In het onderzoek werd ook duidelijk dat de tijd tussen de toetsen ook van belang is. Een belangrijke voorwaarde voor een positief effect op toetsresultaten lijkt te zijn dat de tussentijdse toetsen niet vrijblijvend zijn maar onderdeel uitmaken van de reguliere lessen of op de een of andere manier meetellen in de (cumulatieve) eindscore (zie ook Foss & Pirozello, 2017). Hiermee samenhangend is de rol van de docent dus heel belangrijk bij formatief toetsen. Veel onderzoek geeft namelijk aan dat studenten uit zichzelf gebruik maken van suboptimale leerstrategieën zoals samenvatten of herlezen (Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009; Morehead et al., 2016). De docent heeft dus een belangrijke rol in het helpen met het plannen van de tussentijdse toetsmomenten (zie ook Heitink, van der Kleij, Veldkamp, Schildkamp & Kippers, 2016). Door de toetsing in te bedden in het curriculum, maken de tussentijdse toetsmomenten (i.e., leermomenten) expliciet onderdeel uit van het onderwijs en kan de docent sturen in het aantal toetsmomenten en de spreiding ertussen. Daarnaast laat onderzoek zien dat toetsing met digitale middelen positieve effecten heeft op self-efficacy, betrokkenheid bij de lessen, aanwezigheid, en waargenomen kwaliteit van de lessen (Hunsu, Adesope & Bayly, 2016; Mayer et al., 2009). Een bijkomend voordeel is dat de toetsen met behulp van digitale middelen

mogelijk helpen bij het plannen van de toetsmomenten waardoor tussentijds toetsen meer zelfgestuurd kunnen worden ingezet (Ariel & Karpicke, 2018). Tevens heeft het gezamenlijk beantwoorden van de toetsvragen het effect dat er meer informatie opgehaald wordt waardoor het effect op het leren groter is (Stenlund et al., 2017). Echter, de rol van de student zelf moet niet onderschat worden. Onder andere Fletcher en Shaw (2012) laten zien dat leerlinggestuurde toetsing waarin leerlingen o.a. hun eigen doelen kunnen kiezen en de manier waarop zijn aantonen deze behaald te hebben, leidt tot meer betekenisvol leren. Maar, een leerling of student moet wel in staat zijn goed in te kunnen schatten waar hij staat in het leerproces. Docent gestuurd tussentijds toetsen kan juist helpen om dit inzicht te verkrijgen en vanuit daar vervolgstappen te bepalen (Kang, 2010; Little & McDaniel, 2015; Yeo & Fazio, 2018). Deze overwegingen over de belangrijke rol van de docent bij de inzet van formatieve toetsen leidt tot een tiende richtlijn voor formatief toetsen:

10. Bed formatieve toetsen bewust in het toetsprogramma in, waarbij de programmering geen vrijblijvend maar sturend karakter heeft.

Naast het feit dat deze overzichtsstudie richtlijnen voor formatief toetsontwerp heeft opgeleverd, heeft het uitvoeren ervan ook laten zien dat er knelpunten zijn in het onderzoek naar formatief toetsen. Een deel van deze knelpunten zijn dezelfde als ook al aangeduid door Sluijsmans et al. (2013, p. 71), zoals conceptuele verwarring in de literatuur rondom het begrip formatieve toetsing (wat verstaan we onder formatief toetsen?) en de kwaliteit van de studies (zijn onderzoeksresultaten wel valide, betrouwbaar, bruikbaar en vooral generaliseerbaar?). We hebben daarom geprobeerd voor deze overzichtsstudie het begrip ‘formatieve toets’ zo helder mogelijk af te bakenen in de inleiding van dit rapport. Deze afbakening was niet bedoeld om het gebruik van het concept formatief toetsen in zijn algemeenheid ter discussie te stellen maar had puur een praktisch aspect, aangezien de geheugenpsychologie zich voornamelijk richt op deze smalle vorm van formatieve toetsen. Adesope et al. (2017) en Roediger en Butler (2011) bespreken echter ook andere vormen van tussentijds ophalen van informatie, bijvoorbeeld via flashcards, via het zelf bedenken van vragen of het gebruiken van een korte pauze (3-5 sec) na het stellen van een vraag in de klas. Deze vormen komen overeen met vormen van formatief

toetsen die ook genoemd worden in de literatuur rondom formatieve toetsing (Sluijsmans et al., 2013). Een bijkomend aspect door de focus op de literatuur uit de geheugenpsychologie is dat de meeste studies uitgevoerd zijn in lab settings. Echter, de laatste jaren is er ook steeds meer onderzoek gedaan met onderwijs relevant leermateriaal, langere retentie intervallen en met leeruitkomsten die beter aansluiten op dat wat van leerlingen verwacht wordt in het onderwijs (begrijpen, toepassen). De overzichtsstudie van Adesope et al. (2017) laat zien dat juist deze 'classroom' studies met authentieke toetsen de grootste effecten van retrieval practice laten zien. Mogelijke verklaringen hiervoor zijn meer retentie op de tussentijdse toetsen (doordat er meer geleerd is voorafgaand aan de toets), meer gebruik van hybride toetsen en grotere retentie intervallen (zie ook Son & Simon, 2012). Het lijkt dus dat in elk geval in de *retrieval practice* en *distributed practice* literatuur wel steeds meer aandacht komt voor de praktische relevantie van de bevindingen voor het onderwijs en er langzaamaan een verschuiving van practice tests naar formative tests zichtbaar wordt (zie bijv. de discussie van Adesope et al., 2017). Een bijzondere uitdaging voor meer onderwijs-relevant onderzoek naar formatief toetsen zal echter bestaan uit het creëren van een goede controlegroep om zo de meerwaarde van tussentijds toetsen versus andere methoden zoals samenvatten, concept mappen etcetera te kunnen bepalen (zie ook Moreira, Pinto, Starling, & Jaeger, 2019). Ook zou er meer onderzoek gedaan moeten worden naar transfer (het toepassen van kennis bij nieuwe vragen en in nieuwe leerstof) en hogere orde cognitieve vaardigheden (zie Agarwal, 2018) om zo de relevantie van het onderzoek voor de praktijk te vergroten. Daarmee samenhangend zou er meer aandacht moeten zijn voor de indirecte effecten zoals het effect van tussentijds toetsen op het monitoren en sturen van het leerproces (zie Adesope et al., 2017; Fletcher & Shaw, 2012; Son & Simon, 2012). Als laatste zou in de literatuur naar formatieve toetsen meer aandacht geschonken kunnen worden aan het geheugeneffect van tussentijds toetsen en wat we al weten over effectieve ontwerprichtlijnen (zie Day, Blankenstein, Westenberg & Admiraal, 2017). Juist in de formatieve toetsliteratuur lijkt hier (nog) geen aandacht voor te zijn.

Concluderend, ons doel was om concrete richtlijnen te formuleren voor het ontwerpen van formatief toetsen op basis van bewezen effecten uit de geheugenpsychologie en daarmee bij te dragen aan een sterkere theoretisch onderbouwing voor de effecten van formatief toetsen. De concrete richtlijnen met daarbij een empirische en theoretische onderbouwing zijn in dit rapport geformuleerd. De toegevoegde waarde van deze review bovenop recente reviews

zoals Adesope et al. (2017) en Moreira et al. (2019) is de koppeling van de literatuur over *retrieval practice* en *distributed practice* aan formatieve toetsing en de presentatie van daaruit afgeleide concrete ontwerprichtlijnen voor formatieve toetsen. Wij menen dat met de ontwerprichtlijnen formatief toetsen nog beter als leer- en instructiestrategie kan worden ingezet. Door zowel de directe (geheugeneffect) en indirecte effecten (inzicht in en sturing van het leerproces) van formatieve toetsen te erkennen, is er - theoretisch en praktisch - een grote winst in de onderwijspraktijk te behalen.

Referenties

*Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701. doi:10.3102/00346543166893.

*Agarwal, P. K. (2018). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189-209. doi:10.1037/edu0000282.

Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: A review of publications in French. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 241–264). Paris, France: Organisation for Economic Cooperation and Development.

Assessment Reform Group (2002). *Assessment for learning: 10 principles. Research-based principles to guide classroom practice.* Opgehaald via http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Afl_principles.pdf

*Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*(1), 43-56. http://dx.doi.org/10.1037/xap0000133

*Bae, C. L., Therriault, D. J., & Redifer, J. L. (2018). Investigating the testing effect: retrieval as a characteristic of effective study strategies. *Learning and Instruction.* Advance online publication. https://doi.org/10.1016/j.learninstruc.2017.12.008

*Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extent literature. In J. S. Nairne (Ed.), *The foundations of remembering: essays in honor of Henry L. Roediger III* (pp. 83–105). London: Psychology Press.

*Bangert–Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89–99. doi:10.1080/00220 671.1991.10702818.

*Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology, 32*(2), 201-211. doi:10.1080/01443410.2011.638619

Bartoszweski, B. L., & Gurung, R. A. R. (2015). Comparing the relationship of learning strategies and exam score. *Scholarship of Teaching and Learning in Psychology, 1*(3), 219-228. http://dx.doi.org/10.1037/stl0000036.

*Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. doi: 10.1080/0969594X.2010.513678.

*Bennett, R. E., & Gitomer, D. H. (2009). *Transforming K-12 assessment: integrating accountability testing, formative assessment, and professional support.* Research Memorandum. New York: Princeton, ETS.

*Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *32*(2), 437-452.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principle, Policy, & Practice, 5*(1), 7-74. http://dx.doi.org/10.1080/0969595980050102.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Bjork, R. A., & Bjork, E. L., (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive "processes: Essays in honor of William K. Estes* (vol. 2, pp. 35-67). Hillsdale, New York: Erlbaum.

*Bridger, E. K., & Mecklinger, A. (2014). Errorful and errorless learning: The impact of cue–target constraint in learning from errors. *Memory & Cognition, 42*(6), 898-911.

Brookhart, S. (2001). Successful students' formative and summative used of assessment information. *Assessment in Education: Principles, Policy, and Practice, 8*(2), 153-169.

Brookhart, S. M. (2007). Expanding views about formative assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into Practice* (pp. 43-62). New York: Teachers College Press.

*Butler, A. C., Marsch, E. J., Goode, M. K., and Roediger, H. L. III (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941-956.

*Butler, A. C., & Roediger, H. L. III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

*Cannella-Malone, H. I., Axe, J. B., & Parker, E. D. (2009). Interteach preparation: A comparison of the effects of answering versus generating study guide questions on quiz scores. *Journal of The Scholarship of Teaching and Learning, 9*(2), 22-35.

*Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition, 43*(2), 193-205.

*Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. http://dx.doi.org/10.1037/a0017021.

*Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current directions in Psychological Science, 21*(5), 279-283.

*Carpenter, S., Cepeda, N., Rohrer, D., Kang, S., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369-378. doi:10.1007/s10648-012-9205-z

*Carpenter, S. K. & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619-636.

*Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing *distributed practice:* theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236-246. doi: 10.1027/1618-3169.56.4.236.

*Cepeda, Pashler, Vul, Wixted, & Rohrer, (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychology Bulletin, 132*(3), 354-380.

*Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095-1102.

*Chan, J. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153-170. doi:10.1016/j.jml.2009.04.004

*Chan, J. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49-57. doi:10.1080/09658210903405737

*Chase, J. A., & Houmanfar, R. (2009). The differential effects of elaborate feedback and basis feedback on student performance in a modified, personalized system of instruction course. *Journal of Behavioral Education, 18*, 245-265.

Clark, M. (2012). *What matters most for student assessment systems: a framework paper.* Washington: the World Bank. Opgehaald via https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350wp00publ0wp10read0web04019012.pdf?sequence=1.

*Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition, 40*(4), 528-539

*Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*(6), 919-940. doi: 10.1080/09541440802413505.

*Day, I. Z., Blankenstein, F. v., Westenberg, M., & Admiraal, W. (2018). A review of the characteristics of intermediate assessment and their relationship with student grades. *Assessment & Evaluation In Higher Education, 43*(6), 908-929. doi:10.1080/02602938.2017.1417974.

Delaney, P. F., Verkoeijen, P.P.J.L. & Spirgel, A.(2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of Learning and Motivation-Advances in Research and Theory, 53, 63-148.

Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). *Putting the testing-effect to the test. Why and when is testing effective for learning in secondary school?* Proefschrift. Heerlen: Open Universiteit Nederland.

Dochy, F., Segers, M., Gijbels, D., & Struyven, K. (2007). Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In D. Boud & N. Falchikov (Eds), *Rethinking assessment in higher education: Learning for the longer term* (pp. 87-100). Oxford: Routledge.

*Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: now you see it, now you don't. *Journal of Applied Psychology, 84,* 795–805.

*Dunlosky, J., Rawson, Marsh, Nathan & Willingham (2013). Strengthening the student toolbox: study strategies to boost learning. *American Educator, 37*(3), 12–21;.

*Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paidéia*, *23*(56), 397-406. doi:10.1590/1982-43272356201314

*Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717-741. doi:10.1007/s10648-015-9348-9

*Fitch, M. L., Drucker, A. J., & Norton, J. A. (1951). Frequent testing as a motivating factor in large lecture classes. *The Journal of Educational Psychology, 42*(1), 1-20.

*Fletcher, A. & Shaw, G. (2012). How does student-directed assessment affect learning? Using assessment as a learning process. *International Journal of Multiple Research Approaches, 6*(3), 245-263.

*Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, *109*(8), 1067-1083. doi:10.1037/edu0000197

*Galvagno, S. J., & Segal, B. S. (2009). Critical action procedures testing: A novel method for test-enhanced learning. *Medical Education*, *43*(12), 1182-1187. doi:10.1111/j.1365-2923.2009.03533.x

*Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition and education. *Trends in Neuroscience and Education, 4*(3), 49-59.

*Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31. Opgehaald via http://eprints.glos.ac.uk/3609/1/LATHE%201.%20Conditions%20Under%20Which%20Assessment%20Supports%20Students%27%20Learning%20Gibbs_Simpson.pdf

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*, 95-112.

Goossens , N. A. M. C., Camp, G., Verkoeijen, P. J. L., Tabbers, H. K., & Zwaan, R. A. (2012): Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology, 24*(8), 965-971. doi: 10.1080/20445911.2012.722617.

*Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333-2351. doi:10.1016/j.compedu.2011.06.004.

*Harrington, M., & Jiang, W. (2013). Focus on the forms: recognition practice in Chinese vocabulary learning. *Australian Review Of Applied Linguistics*, *36*(2), 132-145.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134. doi: 10.3758/s13423-011-0181-y.

*Heitink, M. C., van der Kleij, F. M., Veldkamp, B. P, Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17*, 50-62.

*Hinze, S. R. & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 31–40.

*Hung, H. (2015). Intentional Vocabulary Learning Using Digital Flashcards. *English Language Teaching, 8*(10), 107-112.

*Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education, 94*, 102–119. doi:10.1016/j.compedu.2015.11.013.

*Hupbach, A. (2015). *Retrieval practice* does not safeguard memories from interference-based forgetting. *Learning and Motivation, 4,* 23-30.

*Jain, V., Agarwal, V., & Biswas, S. (2012). Use of formative assessment as an educational tool. *Journal of Ayub Medical College Abbottabad, 24*(3

*Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *Plos ONE*, *8*(8), 1-9. doi:10.1371/journal.pone.0070270.

*Johnson, D. I., & Mrowka, K. (2010). Generative learning, quizzing and cognitive learning: An experimental study in the communication classroom. *Communication Education*, *59*(2), 107-123. doi:10.1080/03634520903524739

*Kang, S. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*(8), 1009-1017. doi:10.3758/MC.38.8.1009

*Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences, 3*, 12–19;

*Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term; should spacing be expanding or equal interval? *Psychonomic Bulletin review, 21*(6), 1544-1550. doi: 10.3758/s13423-014-0636-z.

*Kang, S. H. K., McDermott, K. B., & Roediger, H. III (2007). Test format and corrective feedback modify the effect of testing on long term retention. *European Journal of Cognitive Psychology, 19*(4), 528-558.

*Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. In J. H. Byrne, J. T. Wixted (eds.), *Cognitive psychology of memory. Learning and memory: a comprehensive reference, Vol. 2*, 487–514. Oxford: Academic Press. doi: 10.1016/B978-0-12-809324-5.21055-9.

*Karpicke, J. D. & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27,* 317-326.

Karpicke, J. D., Butler, A.C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471-479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval promotes short-term retention, but equal interval retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704 –719.

*Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching Of Psychology, 42*(2), 174-178. doi:10.1177/0098628315573144

*Keene, A. B., Shiloh, A. L., Dudaie, R., Eisen, L. A., & Savel, R. H. (2012). Online testing from Google Docs™ to enhance teaching of core topics in critical care: A pilot study. *Medical Teacher*, *34*(12), 1075-1077. doi:10.3109/0142159X.2012.716553

*Kerdijk, W., Cohen-Schotanus, J., Mulder, B. F., Muntinghe, F. L. H., & Tio, R. A. (2015). Cumulative versus end-of-course assessment: effects on self-study time and test performance. *Medical Education, 49*(7), 709-716. https://doi.org/10.1111/medu.12756 Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology*, *42*(2), 174-178. doi:10.1177/0098628315573144

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

*Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect in skills learning might last 6 months. *Advances in Health Science Education, 15*, 395-401.

*Kuo, T., & Simon, A. (2009). How many tests do we really need? *College Teaching*, *57*(3), 156-160.

*Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory, 20*(1), 37-47. doi: 10.1080/09658211.2011.631550.

*Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science, 42*(3), 373-388. doi:10.1007/s11251-013-9285-2

*Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: The optimal distribution of learning sessions depends on retention interval. *Memory & Cognition, 42*(5), 729-741.

*Kwan, F. (2011). Formative assessment: the one minute paper versus the daily quiz. *Journal of Instructional Pedagogies, 5*(1), 1-8. Opgehaald via https://files.eric.ed.gov/fulltext/EJ1096979.pdf.

Liberati , A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M, Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-Analyses of studies that evaluate health care interventions: explanation and elaboration. *PLOS Medicine*. https://doi.org/10.1371/journal.pmed.1000100

*Little, J. (2018). The role of multiple-choice tests in increasing access to difficult-to-retrieve information. *Journal of Cognitive Psychology, 30*(6), 520-531. https://doi.org/10.1080/20445911.2018.1492581.

*Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following *retrieval practice* for text. *Memory & Cognition, 43*, 85-98.

*Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory, 19*(4), 346-359.

*Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning, 7*(3), 175-195. doi:10.1007/s11409-012-9090-3

*Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory, 20*(8), 899-906.

*Martin, A. J., & Lazendic, G. (2018). Achievement in large-scale national numeracy assessment: An ecological study of motivation and student, home, and school predictors. *Journal of Educational Psychology*, *110*(4), 465-482. doi:10.1037/edu0000231

*Mayer, R. E., Stull, A., DeLeeuw, K., Ameroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 3,* 51–57. doi:10.1016/j.cedpsych.2008.04.002;

*McDaniel, M. A. Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. III (2013). Quizzing in middle-school science; successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27,* 360-372.

*Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General, 145*(7), 897.

*Motz, B. A., de Leeuw, J. R., Carvalho, P. F., Liang, K. L., & Goldstone, R. L. (2017). A dissociation between engagement and learning: Enthusiastic instructions fail to reliably improve performance on a memory task. *PloS one, 12*(7), e0181775.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2015). Instructor and student knowledge of study strategies. *Memory*, 24(2), 1–15. https://doi.org/10.1080/09658211.2014.1001992.

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: a review of applied research. *Frontiers in education, 4(5)*. doi: 10.3389/feduc.2019.00005.

*Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 859-871.

*Narloch, R., Garbin, C. P., & Turnage, . D. (2006). Benefits of prelecture quizzes. *Teaching of Psychology, 33*(2), 109-112.

*Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: research at the interface between cognitive science and education. In R. Scott and S. Kosslyn (Eds), *Emerging Trends in the Social and Behavioral Sciences.* New Jersey: John Wiley & Sons inc.

*Pan. S. C., & Agarwal, P. K. (2018). *Retrieval practice and transfer of learning. Fostering students' application of knowledge.* UCS: San Diego

*Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, doi:10.1037/bul0000151

*Park, J, & Choi, B. C. (2008). Higher retention after a new take-home computerized test. *British Journal of Educational Technology, 39,* 538-545.

*Phelps, R. P. (2012). The effect of testing on student achievement, 1910-2010. *International Journal of Testing, 12,* 21-43.

*Putnam, A. L., & Roediger, H. I. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36-48. doi: 10.3758/s13421-012-0245-x

*Pyc and Rawson (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25*(1), 87-95.https://doi.org/10.1002/acp.1646

Rawson, K. A. (2015). The status of the testing effect for complex materials. Still a winner. *Educational Psychology Review, 27,* 327-331.

*Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: improving performance on course exams and long-term retention. *Educational Psychology Review, 25*, 523-548.

*Roediger, H. I., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends In Cognitive Sciences, 15*(1), 20-27. doi:10.1016/j.tics.2010.09.003

*Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745–6916.2006.00012.x

* Roediger, H. L. III., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology, 31*(5), 1155-1159.

*Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mester & B. Ross (Eds.), *The psychology of learning and motivation: cognition in education* (pp. 1–36). Oxford: Elsevier.

*Rohrer, D. (2009). Avoidance of overlearning characterises the spacing effect. *European Journal of Cognitive Psychology, 21*(7), 1001-1012.

*Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psyhcology: Learning, Memory, and Cognition, 36*(1), 233-239.

*Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing-effect. *Psychological Bulletin, 140*(6), 1432-1463. SNEEUWBAL

Sadler, R. D. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18, 119-144

*Sayeski, K. L., Earle, G. A., Eslinger, R. P., & Whitenton, J. N. (2017). Teacher candidates' mastery of phoneme-grapheme correspondence: massed versus distributed practice in teacher education. *Annals of dyslexia, 67(*1), 26-41.

*Schellenberg, S., Negishi, M., & Eggen, P. (2011). The effects of metacognition and concrete encoding strategies on depth of understanding in educational psychology. *Teaching Educational Psychology, 7*(2), 17-24.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

*Sluijsmans, D., Joosten-ten Brinke, D., & van der Vleuten, C. (2013). *Toetsen met leerwaarde. Een reviewstudie naar de effectieve kenmerken van formatief toetsen*. Den Haag:

NWO-PROO Opgehaald via https://www.nro.nl/wp-content/uploads/2014/05/PROO+Toetsen+met+leerwaarde+Dominique+Sluijsmans+ea.pdf.

*Smith, M. A., & Karpicke, J. D. (2014) Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784-802. doi: 10.1080/09658211.2013.831454

*Son, L., & Simon, D. (2012). Distributed learning: data, metacognition, and educational implications. *Educational Psychology Review, 24*(3), 379-399. doi:10.1007/s10648-012-9206-y

*Stenlund, T., Sundström, A., & Jonsson, B. (2016) Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology, 36*(10), 1710-1727. doi: 10.1080/01443410.2014.953037

*Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany, NY: State University of New York Press.

*Taveira Gomes, T., Prado-Costa, R., Severo, M., & Ferreira, M. A. (2015). Characterization on medical students recall of factual knowledge using learning objects and repeated testing in a novel e-learning system. *BMC Medical Education, 14*(4).

*Tse, C & Pu, X. (2012). The effectiveness of test enhanced learning depends on trait test anxiety and working memory capacity. *Journal of experimental psychology: applied, 18*(3), 253-264.

*Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in a large, diverse introductory college psychology course. *Journal of Experimental Psychology: Applied, 22*(2), 148.

*Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal Of Experimental Psychology: General, 143*(4), 1526-1540. doi:10.1037/a0036036

*Van Gog, T., & Kester, L., (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36,* 1532-1541.

*Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increased. *Educational Psychological Review, 27,* 247-264.

*Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & Cognition, 44*(6), 897-909.

Van der Vleuten, C. P. M., & Driessen, E. W. (2000). *Toetsing in probleemgestuurd onderwijs*. Wolters Noordhoff, Groningen.

Verdaasdonk, E. G. G., Stassen, L. P. S., van Wijk, R. P. J., & Dankelman, J. (2007). The influence of different training schedules on the learning of psychomotor skills for endscopic surgery. *Surgical Endoscopy, 21*(2), 214-219. https://doi.org/10.1007/s00464-005-0852-8.

*Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development, 83*(4), 1137-1144.

*Vojdanska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: the role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology, 24,* 1183-1195.

*Volante, L. (2010). Assessment of, for, and as learning within schools: Implications for transforming classroom practice. *Action in Teacher Education, 31*(4), 66-75.

*Weinstein, Y., Sumeracki, M. A., & Madan, C. R. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications, 3*(2). https://doi.org/10.1186/s41235-017-0087-y

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1036-1046.

Wiliam, D. & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for k-12 classrooms.* West Palm Beach, Florida: Learning Sciences International.

*Yeo, D. J., & Fazio, L. K. (2018). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology.* Advance online publication. http://dx.doi.org/10.1037/edu0000268.

*Yiğit, A., Kıyıcı, F. & Çetinkaya, G. (2014). Evaluating the testing effect in the classroom: An effective way to retrieve learned information. *Eurasian Journal of Educational Research, 54,* 99-116.