# Prestaties op papieren en digitale examens: wat is het verschil?

Verslag van een literatuurstudie

Eindrapportage mei 2015

***Hans Kuhlemeier***

***Hendrik Straat***

***Paul van der Molen***

# Inhoudsopgave



 

# 1 Inleiding

De laatste decennia worden de papier-en-pen-examens (PPT) in toenemende mate vervangen door computer-based examens (CBT). CBT heeft vele voordelen boven PPT, zoals een vermindering van de correctielast, objectivering van de beoordeling, voorkomen van afkijken door toewijzing van verschillend ogende maar gelijkwaardige vraagversies, adaptieve toetsing, mogelijkheid tot het geven van feedback aan kandidaten tijdens de afname, besparing van druk- en verzendkosten en snellere rapportage van de resultaten. Een bijkomend voordeel is dat leerlingen over het algemeen een voorkeur hebben en meer gemotiveerd zijn voor digitale toetsing dan voor papieren toetsing (o.a. Glassnapp, Poggio, Poggio, & Yang, 2005; O'Malley et al., 2005; Ito & Sykes, 2004; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait & Threlfall, 2004).

Naast voordelen worden er ook nadelen genoemd. Al vanaf het begin van digitale toetsing en examinering is de vergelijkbaarheid van de scores op PPT en CBT een punt van zorg geweest (o.a. Green, Bock, Humphreys, Linn & Reckase, 1984; APA, 1986; Mazzeo & Harvey, 1988). Een veel genoemd probleem is dat PPT- en CBT-versies van hetzelfde examen tot verschillende resultaten kunnen leiden. Deze zogeheten moduseffecten (Engels: mode effects) verwijzen naar verschillen in examenprestaties ten gevolge van het aanbieden van het examen op papier dan wel via de computer (o.a. Kolen, 1999; Bennett, 2003; Wang, Jiao, Young, Brooks & Olson, 2007). Zo zou CBT bij het vakgebied rekenen-wiskunde gemiddeld tot hogere of lagere prestaties kunnen leiden dan PPT. Lagere prestaties op een CBT kunnen een probleem vormen als de resultaten veel lager zijn dan vrijwel iedereen verwacht en de afnamemodus als een mogelijke oorzaak wordt genoemd. Zo signaleerde de Commissie Bosker (2014) dat de prestaties van de kandidaten op de Rekentoetsen vo en rekenexamens mbo tegenvallen. Bij deze rekentoetsen zou de digitale afname tot gemiddeld lagere prestaties hebben geleid dan wanneer deze toets als PPT was afgenomen. Als mogelijke oorzaken voor de tegenvallende prestaties (die samenhangen met afnamemodus) noemt de Commissie Bosker:

- in de digitale rekentoets ontbreken opgaven die werken met *partial credit* (het belonen van tussenstappen), waardoor één fout in één van de stappen in het oplossen van een opgave ertoe leidt dat nul punten voor het antwoord worden gegeven (waardoor de prestaties op de digitale rekentoets lager zijn dan wanneer de antwoorden beoordeeld zouden zijn via de gebruikelijke *partial* credit scoring van op papier gegeven antwoorden);
- anders dan bij een papieren toets konden de leerlingen tot en met 2014 in de digitale rekentoetsen niet terugbladeren en hun antwoorden bekijken en vervolgens verbeteren (terwijl leerlingen geleerd is om lastige opgaven even te laten liggen); enquêtes onder door docenten en leerlingen bevestigden dat dit als een ernstig knelpunt werd ervaren (College voor Examens, 2014);
- bij de digitale afname van de rekentoets zijn leerlingen minder geneigd kladpapier te gebruiken dan bij afname op papier, wat de kans op fouten vergroot;
- de leerlingen die de digitale rekentoet maken, zijn over het algemeen nog onbekend met het digitaal afnemen van toetsen/examens.

De commissie doet onder meer de volgende aanbevelingen:

- onderzoek de mogelijkheid van schriftelijke afname naast toetsing per computer;
- onderzoek de mogelijkheden om in de rekentoetsen te werken met *partial credit* (het belonen van tussenstappen).

Beide aanbevelingen zijn beleidsmatig van belang. Als PPT- en CBT-versies van dezelfde rekentoets naast elkaar gebruikt zouden worden, vereisen professionele standaarden (APA, 1986; AERA, 1999; International Test Commission, 2001, 2006; Evers, Lucassen, Meijer & Sijtsma, 2009) dat de gerapporteerde scores aantoonbaar vergelijkbaar zijn (Wang & Kolen, 2001). Een duaal afnamesysteem verhoogt de correctielast voor de docenten/beoordelaars, en dat geldt ook als schriftelijke en digitale afname binnen hetzelfde examen gecombineerd zou worden (tenzij de correctie natuurlijk 'uitbesteed' wordt, maar dat lijkt in het Nederlandse examensysteem vooralsnog onwaarschijnlijk). En als de huidige dichotome scoring van alleen het eindantwoord bijvoorbeeld vervangen zou worden door stapsgewijze bevraging met geautomatiseerde toekenning van deelscores, dreigt het gevaar dat nieuw verzamelde gegevens niet meer vergelijkbaar zijn met eerdere gegevens (o.a. Bennett, Braswell, Oranje, Sandene, Kaplan & Yan, 2008). Uiteraard is dit een tijdelijk probleem dat opgelost kan worden, maar het kan wel betekenen dat het referentiekader voor de interpretatie van de scores op de rekentoetsen opnieuw vastgesteld moet worden (College voor Toetsen en Examens, 2014; Béguin & Wools, 2015).

In deze literatuurstudie gaan we na wat er in de Engelstalige literatuur bekend is over het effect van de afnamemodus op de leerprestaties. Omdat het commentaar van de Commissie Bosker op de rekentoetsen de aanleiding vormde voor deze literatuurstudie, ligt het accent op verschillen tussen CBT en PPT bij het vakgebied rekenen-wiskunde.
Allereerst kijken we naar wat er in algemene zin bekend is over het effect van de afnamemodus op de hoogte van de prestaties. In het kader van een kortdurende studie is het niet mogelijk om de vele honderden uitgevoerde vergelijkingsstudies te bestuderen en samen te vatten. We beperken ons tot het synthetiseren van de uitkomsten van de belangrijkste reviews en meta-analyses (Mazzeo & Harvey, 1988; Bunderson, Inouye & Olsen, 1989; Wise & Plake, 1989; Bergstrom, 1992; Dillon, 1992; Mead en Drasgow, 1993; Kim, 1999; Russell, Goldberg & O'Connor, 2003; Paek, 2005; Gaskill & Marshall; 2006; Wang, Jiao, Young, Brooks & Olson, 2007; Texas Education Agency, 2008; Kingston, 2009). Daarbij kijken we met name naar het eventuele effect van de afnamemodus op de gemiddelde prestaties van de kandidaten (en niet zozeer naar eventuele effecten op bijvoorbeeld betrouwbaarheid, validiteit, rangordening van de kandidaten, indicatoren van classificatieaccuratesse zoals percentage onvoldoendes en zak-/slaagpercentages, of de stabiliteit van de itemparameters).
Vervolgens gaan we op zoek naar wat er bekend is over mogelijke oorzaken van het effect van de afnamemodus op de prestaties. Kolen (1999-2000) noemt vier clusters van verklarende factoren die ertoe kunnen bijdragen dat kandidaten hoger of juist lager scoren op de ene afnamemodus in vergelijking met de andere: a) verschillen in de examenvragen, b) verschillen in de scoring van de antwoorden, c) verschillen in de afnamecondities en d) verschillen in de geëxamineerde groep kandidaten (zoals naar geslacht, sociaal milieu, etniciteit, vaardigheidsniveau, ICT-ervaring en toetsangst). Bennett (2003) maakt een onderscheid in kenmerken van de itempresentatie (bijvoorbeeld het aantal items per scherm versus per pagina), de vereiste response (bijvoorbeeld het aankruisen van een antwoordalternatief op papier versus het aanwijzen, klikken, slepen, positioneren en scrollen met de muis of het gebruik van het toetsenbord om tekst in te voeren en te wijzigen). In het kader van een kortdurende studie is het niet mogelijk om al deze factoren te onderzoeken. We beperken ons tot verschillen tussen CBT en PPT die de Commissie Bosker genoemd heeft als mogelijke oorzaken voor de tegenvallende prestaties op de Rekentoetsen.

Onderzoeksvragen

Al met al onderscheiden we in deze review de volgende vijf onderzoeksvragen:

- Wat is er bekend over het effect van de afnamemodus op de hoogte van de prestaties van de kandidaten (in algemene zin en bij rekenen-wiskunde in het bijzonder)?
- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met het al dan niet toekennen van deelscores *(partial credit)* aan gedeeltelijk goede antwoorden? Hierbij nemen we ook de vraag mee in hoeverre het tegenwoordig mogelijk is de resultaten van de leerlingen op een CBT-examen op eenzelfde manier te beoordelen als de gebruikelijke *partial credit* beoordeling in een PPT-examen.
- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met de verschillende mogelijkheden tot item review?
- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met verschillen in het gebruik van kladpapier?
- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met de manier waarop kandidaten hun antwoord moeten geven: het intypen van het antwoord in een invoervak op het beeldscherm volgens de specifieke vereisten van de digitale invoermodule versus het noteren van het antwoord op papier volgens de gebruikelijke en vertrouwde wiskundige notatiewijzen en conventies.

- De vijfde onderzoeksvraag wordt niet letterlijk door de Commissie Bosker (2014) genoemd, maar heeft wel raakvlakken met de door de commissie geconstateerde onbekendheid met het digitaal afnemen van toetsen/examens als een van de verklaringen voor de tegenvallende resultaten. Kandidaten hebben vaak meer ervaring met het noteren van antwoorden op wiskundevragen op papier dan met het invoeren van deze antwoorden via het beeldscherm (McGuire & Youngson, 2002). Daardoor zouden kandidaten bij een CBT meer invoerfouten maken dan bij een PPT, met als gevolg lagere prestaties voor CBT in vergelijking met PPT.

Bij de interpretatie van de uitkomsten van deze literatuurstudie is een relativerende kanttekening op zijn plaats. Reeds in de jaren tachtig van de vorige eeuw merkten Bunderson, Inouye en Olsen (1989) op dat digitale afnamesystemen zo snel veranderen dat elke poging om de balans op te maken al snel verouderd zal zijn. Zij voegen daar aan toe dat *"Today's state-of-the-art devices might become exhibits in tomorrow's museum of antiquities*" (p. 21). Dit geldt tot op zekere hoogte ook voor de rekentoetsen die met ingang van 2015 sterk verbeterd zijn terwijl er nog meer verbeteringen in het verschiet liggen. Zo bieden de nieuwe rekentoetsen tegenwoordig ruime mogelijkheden tot item review (dit wil zeggen: het terugbladeren, bekijken en veranderen van antwoorden). Daarnaast laten zeer recente vragenlijstgegevens (Cito, 2015a) zien dat de overgrote meerderheid van ruim 16000 bevraagde kandidaten van mening is dat het computerprogramma waarin de rekentoets staat duidelijk is (96%) en dat het aangeven of intypen van de antwoorden bij de opgaven gemakkelijk is (86%). Verder blijkt uit nog ongepubliceerde gegevens dat syntactische invoerfouten nauwelijks meer voorkomen (zoals het ten onrechte gebruiken van een punt voor het aangeven van duizendtallen). Tot slot wordt er gewerkt aan geautomatiseerde scoringsalgoritmen die de traditionele menselijke beoordeling van antwoorden op papier waar mogelijk en wenselijk proberen na te bootsen. Op basis van een inhoudsanalyse van veel voorkomende fouten en gedeeltelijk goede antwoorden van kandidaten (o.a. Cito, 2015b) is ervaring opgedaan met het waarderen van nu nog als fout gescoorde antwoorden die een menselijke beoordelaar bij een papieren examen waarschijnlijk wel

(gedeeltelijk) goed gerekend zou hebben. Een voorbeeld is het toekennen van punten aan alle ingevoerde geldbedragen die minder dan 10 cent verschillen van het volledig correcte antwoord. In plaats van slechts één goed antwoord te accepteren, rekent het scoringsalgoritme alle antwoorden binnen een bepaald scorebereik goed. Uiteraard is deze versoepeling van de scoring niet bij alle itemtypen mogelijk en wenselijk, maar het geeft wel aan dat de rekentoetsen volop in beweging zijn.

# 2 Methode van onderzoek

De literatuurstudie is uitgevoerd in het begin van 2015. De werkzaamheden zijn uitgevoerd in opdracht van het College voor Toetsen en Examens (CvTE).

Ter beantwoording van de eerste en tweede onderzoeksvraag zijn twee literatuur searches uitgevoerd. Daarbij is gezocht in de databases ERIC, PsycINFO, Applied Social Sciences Index and Abstracts (ASSIA), International Bibliography of the Social Sciences (IBSS), Library and Information Science Abstracts (LISA) en ProQuest Social Science Journals. In aanmerking kwamen publicaties die zijn verschenen in de periode 1970 tot en met 2015. De beslissing om ook minder recente publicaties mee te nemen is welbewust genomen. De rekentoetsen waar de Commissie op doelt, behoren namelijk tot de categorie eerste-generatie CBT (Bunderson, Inouye & Olsen, 1989) uit de beginperiode van het digitale tijdperk. Denk bijvoorbeeld aan het ontbreken van mogelijkheden tot a) item review, b) geautomatiseerde *partial credit scoring,* c) feedback op onder meer de syntactische juistheid van ingetypte antwoorden, d) digitaal markeren en annoteren, e) adaptieve toewijzing van items van een verschillende moeilijkheidsgraad afhankelijk van het vaardigheidsniveau van de kandidaat, en f) moderne itemtypen met gebruikmaking van bijvoorbeeld animaties en simulaties.

Ter beantwoording van de eerste onderzoeksvraag naar het effect van de afnamemodus op de prestaties is in eerste instantie na enig uitproberen de volgende zoekstring gehanteerd: "((compara*) OR (equivalen*)) AND ((computer) or (on-line)) AND ((test*) or (assess*))". Deze zoekactie resulteerde in 70.522 publicaties. Dit aantal was veel te groot om binnen de temporele en budgettaire beperkingen van het project bestudeerd te worden. Daarom is dezelfde zoekstring nogmaals toegepast, maar nu is er alleen in de Abstracts van de publicaties gezocht (in plaats van in de volledige tekst). Dit resulteerde in 2.245 publicaties. Er is besloten de zoekactie te beperken tot de titel van de publicatie. De zoekstring "ti((compara*) OR (equivalen*)) AND ti((computer) or TI(on-line)) AND ti((test*) or (assess*))" leverde 44 publicaties op waarvan acht on-topic.

Ter beantwoording van de tweede onderzoeksvraag naar het toekennen van *partial credit* werd een tweede literatuursearch uitgevoerd. Daarbij werd uiteindelijk de zoekstring "(TI(partial credit) OR ti(partial knowledge)) and TI(math*)" gehanteerd. Dit leverde zeven publicaties op waarvan twee on-topic. Voor de derde, vierde en vijfde onderzoeksvraag is vanwege de beperkte tijd geen afzonderlijke literatuur search uitgevoerd. Het aantal gemiste publicaties zal hier dus groter zijn dan bij de eerste twee onderzoeksvragen.

Naast de beide literatuur searches is een serie Web-searches uitgevoerd met de zoekmachines Google en Google Scholar. De aldus gevonden publicaties zijn gescreend op verwijzingen naar reviews en meta-analyses over het effect van de afnamemodus op de prestaties. Dit leverde in totaal dertien publicaties op (Mazzeo & Harvey, 1988; Bunderson, Inouye & Olsen, 1989; Wise & Plake, 1989; Bergstrom, 1992; Dillon, 1992; Mead en Drasgow, 1993; Kim, 1999; Russell, Goldberg & O'Connor, 2003; Paek, 2005; Gaskill & Marshall, 2006; Wang, Jiao, Young, Brooks & Olson, 2007; Texas Education Agency, 2008; Kingston, 2009). Deze reviews en meta-analyses vormden het uitgangspunt voor de beantwoording van de eerste onderzoeksvraag.
De reviews, meta-analyses en overige publicaties zijn vervolgens doorzocht op relevante verwijzingen naar studies over *partial credit,* item review, kladpapier en invoer op papier versus via beeldscherm.

Dit alles resulteerde in totaal in circa 150 publicaties die voor onze onderzoeksvragen relevant zouden kunnen zijn. Geprobeerd is deze publicaties in het pdf-format te verkrijgen via onder meer het Internet en het Kenniscentrum van Cito, wat in vrijwel alle gevallen gelukt is. Na bestudering bleken niet alle 150 publicaties voor het doel van ons onderzoek bruikbaar. Voor de publicaties die uiteindelijk voor deze literatuurstudie gebruikt zijn, wordt verwezen naar het overzicht van geraadpleegde literatuur aan het einde van dit rapport.

Beperkingen

Naast het feit dat deze literatuurstudie in mei 2015 afgerond diende zijn, moet er bij de interpretatie van de resultaten ook rekening worden gehouden met de volgende beperkingen:

- Bij het zoeken zijn geen eisen gesteld aan de methodologische zuiverheid van de studies. Peer-review vormde bijvoorbeeld geen selectiecriterium.
- De kans dat relevante publicaties gemist zijn, is relatief groot in vergelijking met de brede en diepgaande reviews zoals deze door het Nationaal Regieorgaan Onderwijsonderzoek (NRO) worden uitgezet (zie bijvoorbeeld Scheltinga, Keuning & Kuhlemeier, 2015).
- Het onderzoeksobject is voornamelijk beperkt tot het effect van de afnamemodus op de hoogte van de prestaties. Grotendeels buiten beschouwing blijven hiermee onder meer het effect op de validiteit, betrouwbaarheid, rangordening van de kandidaten, indicatoren van classificatieaccuratesse zoals percentage onvoldoendes en zak-/slaagpercentages, en de stabiliteit van de itemparameters.
- In deze literatuurstudie is vrijwel uitsluitend gekeken naar de vergelijkbaarheid van de totaalscores op PPT en CBT, en slechts af en toe naar verschillen op het niveau van de subtoets of het individuele item. Dit laatste is van belang omdat er sterke moduseffecten op subtoets- of itemniveau kunnen zijn die elkaar uitmiddelen op het niveau het de totaalscore (Pommerich, 2004; Johnson & Green, 2006).
- Met betrekking tot *partial credit* is alleen gekeken naar het mogelijke effect van het al dan niet toekennen van deelscores in de context van CBT. De zeer omvangrijke maar nu enigszins gedateerde literatuur over het effect van dichotome versus polytome scoring van meerkeuze- versus open vragen op papier kon niet worden meegenomen (voor een review, zie Traub, 1993). Ook laten we *partial credit* scoring in het kader van formatieve evaluatie van wiskundige vaardigheden in oefenomgevingen buiten beschouwing (o.a. DWO, 2013).
- In de bespreking van individuele studies beperken we ons voornamelijk tot het rapporteren van de belangrijkste onderzoeksresultaten. Gezien de beperkte tijd bleek het niet haalbaar om de gevonden studies te bespreken volgens het gebruikelijke rapportagestramien met als onderdelen: aanleiding, context, vraagstelling, methode van onderzoek (proefpersonen, design, data-analyse), resultaten, discussie, conclusies en aanbevelingen; zie bijvoorbeeld Scheltinga, Keuning & Kuhlemeier, 2015).

# 3 Resultaten

## 3.1 Het effect van de afnamemodus op de prestaties

**Inleiding**

De eerste onderzoeksvraag luidt: "Wat is er bekend over het effect van de afnamemodus op de hoogte van de prestaties van de kandidaten? We beantwoorden deze vraag aan de hand van de verhalende reviews en meta-analyses van Mazzeo en Harvey (1988), Bunderson, Inouye en Olsen (1989), Wise en Plake (1989), Bergstrom (1992), Dillon (1992), Mead en Drasgow (1993), Kim (1999), Russell, Goldberg en O'Connor (2003), Paek (2005), Gaskill & Marshall, 2006; Wang, Jiao, Young, Brooks en Olson (2007), Texas Education Agency (2008) en Kingston (2009).

**Resultaten reviews en meta-analyses**

**Mazzeo en Harvey (1988)**

Mazzeo and Harvey (1988) analyseerden 38 studies (45 tests) waarin zij de resultaten van CBT- en PPT-versies van tests op het gebied van intelligentie, geschiktheid, persoonlijkheid en prestaties met elkaar vergeleken.

Bij elf studies resulteerde de CBT in hogere scores dan de PPT, bij achttien was er geen verschil en bij vijftien waren de PPT-scores hoger.

De auteurs veronderstelden dat speedtests gevoeliger zijn voor modus-effecten dan power tests. Bij een speedtest gaat het erom zoveel mogelijk items binnen de toegewezen tijd correct te beantwoorden. Bij een power test is er geen tijdlimiet of is de tijdlimiet zo ruim dat vrijwel elke kandidaat de test volledig kan maken. De aanname dat speedtests gevoeliger zijn voor modus-effecten dan power tests formuleerden Mazzeo en Harvey (1988) als volgt**:** "*Any distractions or difficulties introduced by automated test administration, or advantages introduced by ease of responding, will affect the rate at which items are answered* (p. 4)". De afnamemodus - papier versus digitaal - bleek inderdaad van geen belang bij power tests, maar bij speedtests bleek de PPT-versie gemiddeld tot hogere scores te leiden dan de CBT-versie. Als mogelijke verklaring verwijzen de auteurs ernaar dat het lezen van langere teksten op het scherm meer tijd kost dan op papier.

Bij de CBT-versies werden meer vragen overgeslagen dan bij de PPT-versies.

Ook de antwoordmodus (beeldscherm versus papier) bleek een mogelijke oorzaak van de hogere scores op PPT. Onderzoek waarbij de door de computer gelezen antwoorden werden vergeleken met de antwoorden op papier liet zien dat de computer lang niet alle ingevoerde antwoorden correct leest.

Andere factoren die zorgden voor verschillen tussen PPT en CBT waren het gebruik van multi-screen items, grafische weergaves en complexe beeldschermweergaves.

**Bunderson, Inouye en Olsen (1989)**

In hun review analyseerden Bunderson, Inouye en Olsen (1989) 23 studies waarbij de scores op PPT werden vergelijken met die op CBT.

Bij drie studies deden de leerlingen het beter op de CBT, bij elf studies was er geen verschil en bij de overige negen behaalden de leerlingen hogere prestaties op de PPT.

In hun bespreking van de bevindingen wijzen Bunderson et al. (1989) onder meer op de scoringsprocedure als mogelijke bron van onvergelijkbaarheid. De hogere scores op PPT worden

mede veroorzaakt doordat de computer ingevoerde antwoorden van kandidaten ten onrechte als fout interpreteert terwijl een menselijke beoordelaar ze op papier goed of gedeeltelijk goed zou rekenen (zie hiervoor ook paragraaf 3.5).
Een andere mogelijk bron van moduseffecten is dat het ontwerp van de interface het niet toestaat om fouten bij het invoeren van antwoorden onmiddellijk te corrigeren of om eerder ingevoerde antwoorden te bekijken en zo nodig te veranderen, zoals bij een PPT-versie van dezelfde toets uiteraard wel mogelijk is. Overigens zijn de auteurs van mening dat het hier een triviaal ontwerpprobleem betreft aangezien er ten tijde van hun review reeds verschillende goede oplossingen beschikbaar waren (die vanaf 2015 ook in de rekentoetsen geïmplementeerd zijn).

**Wise en Plake (1989)**
De review van Wise en Plake (1989) is vooral van belang voor onze derde vraagstelling naar het vermeende negatieve effect van het ontbreken van mogelijkheden tot item review in eerste-generatie CBT op de prestaties. Vandaar dat we de resultaten ervan bespreken in paragraaf 3.3.

**Dillon (1992)**
De review van Dillon (1992) is voor ons minder interessant omdat het vooral betrekking heeft op verschillen tussen het lezen en begrijpen van grote hoeveelheden informatie op het scherm of papier. Een van de conclusies was dat het lezen van het scherm ongeveer 20% tot 30% langzamer gaat dan van papier, een conclusies die anno 2015 vanwege de grote technische verbeteringen waarschijnlijk grotendeels achterhaald is.

**Bergstrom (1992)**
Bergstrom (1992) analyseerde twintig studies zoals beschreven in acht onderzoeksrapporten waarin de prestaties op PPP werden vergeleken met die op adaptieve computertoetsen (CAT). Ondanks verschillen tussen de studies qua getoetste leerstof, de leeftijd van de proefpersonen, het gebruikte IRT-model en het onderzoekontwerp bleken PPT en CAT over het algemeen tot vergelijkbare prestaties te leiden. Na verwijdering van vijf studies die te sterk bijdroegen aan de heterogeniteit van de effectgroottes, vond zij een gemiddelde effectgrootte van -.002. In de studies waarin een significant gemiddeld verschil gevonden werd, waren de scores op de PPT hoger dan die op de CAT als de PPT als eerste was afgenomen. De auteur doet de aanbeveling nader onderzoek te doen naar scrollen als mogelijke verklaring voor de lagere prestaties op CAT.

**Mead en Drasgow (1993)**
In hun meta-analyse analyseerden Mead en Drasgow (1993) een groot aantal studies naar verschillen tussen papieren en digitale afname van power en speedtests. Voor power tests met een tijdslimiet - die in het onderwijs veel gebruikt worden - bedroeg de effectgrootte van het gemiddeld verschil tussen PPT en CBT .03 met een standaarddeviatie van .15. CBT's zijn gemiddeld dus nauwelijks moeilijker dan PPT's en het gemiddeld prestatieverschil varieert weinig van studie tot studie.
Mead en Drasgow (1993) keken ook naar het effect van de afnamemodus op de rangordening van degenen die getoetst werden. Voor power tests was de gemiddelde correlatie (na correctie voor onbetrouwbaarheid) zeer hoog, te weten .91 over 123 geanalyseerde correlaties, wat er op wijst dat de rangordening van de leerlingen nagenoeg gelijk is.
Voor speed tests was de gemiddelde correlatie veel lager, namelijk .72 over 36 correlaties.
De auteurs vonden geen verschil tussen lineaire (CBT) en adaptieve (CAT) digitale toetsing.

**Kim (1999)**
In een meta-analyse bestudeerde Kim (1999) 51 studies naar de vergelijkbaarheid van scores op PPT, lineaire (CBT) en adaptieve (CAT) digitale toetsen bij een groot aantal doelgroepen en vakgebieden. Het gemiddelde over de 226 bestudeerde effectgroottes voor enerzijds CBT en CAT en anderzijds PPT bedroeg .019 (95%-betrouwbaarheidsinterval -.030 tot .068) en was statistisch niet significant. Echter, het verschil bleek sterk afhankelijk van het type CBT. Voor de vergelijking CBT-PPT bedroeg de gemiddelde effectgrootte .103 (lineaire CBT gemiddeld beter gemaakt dan PPT) en voor de vergelijking CAT-PPT ging het om -.125 (CAT gemiddeld slechter gemaakt dan PPT).
Kim rapporteerde de resultaten ook apart voor verschillende leeftijdsgroepen en vakgebieden. Voor high school leerlingen bleek CBT gemiddeld makkelijker dan PPT, wat de auteur toeschreef aan hun positieve houding ten opzichte van toetsing met de computer en de '*excitement*' die het maken van een digitale toets met zich meebracht.

**Russell, Goldberg en O'Connor (2003)**
In hun review analyseerden Russel, Goldberg en O'Connor (2003) onderzoek naar de vergelijkbaarheid van PPT- en CBT-toetsen op het gebied van schrijfvaardigheid en wiskunde. Zij maakten een onderscheid tussen onderzoek in de periode tot 1986 en de periode daarna. Hun review brengt een aantal factoren aan het licht die van invloed zijn op de hoogte van de prestaties (en de validiteit van de meting). Daartoe behoren moduseffecten die te maken hebben met de mogelijkheden om individuele items over te slaan, opnieuw te bekijken en te veranderen en de transfer van informatie tussen scherm en kladpapier. Zie hiervoor ook paragraaf 3.3 en 3.5.
Andere moduseffecten die de auteurs de revue laten passeren, betreffen de presentatie van *graphics* en tekst op het scherm in relatie tot de ervaring van leerlingen met het werken met computers. Omdat deze mogelijke verklaringen voor het doel van onze literatuurstudie van minder groot belang zijn, laten we bespreking ervan hier verder achterwege.

**Paek (2005)**
Paek (2005) bestudeerde een aantal reviews en onderzoeken naar de vergelijkbaarheid van PPT en CBT die vóór en na 1993 waren uitgevoerd. De studies betroffen de vakgebieden rekenen-wiskunde, taal, lezen en science. Zij constateert dat de zogeheten *electronic pageturners* - in het Nederlandse taalgebied ook wel aangeduid als 'rechtop gezette' papieren toetsen - regelmatig slechter gemaakt werden dan de papieren versies. Een eerste verklaring suggereert dat leerlingen moesten wennen aan de nieuwe afnamemodus: kandidaten moesten onder *high-stakes* condities leren te navigeren in een interface waarmee zij nog niet vertrouwd waren. Als tweede verklaring noemt Paek (2005) dat het destijds nog niet mogelijk was om vooruit en achteruit te bladeren, items over te slaan en itemantwoorden te bekijken en te veranderen, wat in PPT en meer recente CBT natuurlijk wel mogelijk is (en wat ten goede komt aan de motivatie van de leerling).

Bij de na 1993 uitgevoerde studies waren de verschillen tussen PPT en CBT kleiner dan bij het daarvoor uitgevoerde onderzoek. In de meeste recentelijk uitgevoerde studies bij zogeheten K12-leerlingen (dit wil zeggen: leerlingen van *kindergarten* tot en met *grade 12*; ofwel van 4 à 6 tot 17 à 19 jaar) bleken de scores op CBT gelijkwaardig of iets hoger dan die op PPT. Waar er verschillen tussen afnamemodi gevonden werden, waren deze lang niet altijd statistisch significant. Paek schrijft deze ogenschijnlijke afname van de verschillen tussen PPT en CBT toe aan de toegenomen ervaring

met computers en het gegeven dat recentere CBT de mogelijkheden tot het navigeren, overslaan en wijzigen van antwoorden van de reguliere PPT beter kunnen nabootsen.

Paek rapporteert de recentere resultaten ook per vakgebied. Zo behaalden de leerlingen bij drie van de twaalf geanalyseerde studies op het gebied van rekenen-wiskunde lagere scores op de CBT, bij één was de PPT moeilijker en bij de overige acht waren de scores vergelijkbaar.

**Gaskill en Marshall (2006)**
De review van Gaskill en Marshall (2006) is voor ons minder interessant omdat het vooral gericht is op het synthetiseren van onderzoek naar interacties tussen de afnamemodus en achtergrondkenmerken van leerlingen, zoals vaardigheidsniveau, geslacht en sociaal-etnische achtergrond. Van groter belang voor ons doel is het onderzoek dat zij zelf uitvoerden naar de vergelijkbaarheid van PPT en CBT bij meerkeuzetoetsen *numeracy* in grade 4 en 7. Omdat dit onderzoek relevant is voor de beantwoording van onze vierde onderzoeksvraag over de rol van kladpapier, verwijzen wij voor de resultaten ervan naar paragraaf 3.4.

**Wang, Jiao, Young, Brooks en Olson (2007)**
In hun meta-analyse analyseerden Wang, Jiao, Young, Brooks en Olson (2007) 38 studies op het gebied van rekenen-wiskunde bij K12-leerlingen. Zij vonden geen significante verschillen qua gemiddelde prestaties tussen PPT en CBT. Dit resultaat werd echter verkregen na verwijdering van zes outlier studies waarin CBT aanzienlijk slechter gemaakt waren dan PPT.
Bennett, Braswell, Oranje, Sandene, Kaplan & Yan (2008) merken op dat de meeste studies uit de meta-analyse van Wang niet officieel gepubliceerd waren (en dus waarschijnlijk niet door vakgenoten beoordeeld zijn). Daarnaast ging het meestal alleen om meerkeuze-opgaven, was de meerderheid van de effectgroottes van slechts drie onderzoeken afkomstig en hielden de onderzoekers geen rekening met de representativiteit van de steekproeven.

**Texas Education Agency (2008)**
Naar aanleiding van een review van uitgevoerde vergelijkingsstudies over een periode van twintig jaar concludeert de Texas Education Agency (2008) dat de onderzoeksresultaten sterk variëren al naar gelang het vakgebied (lezen, rekenen-wiskunde, science et cetera), het itemtype (meerkeuze versus open vragen) en het testprogramma (state assessment programs, toetsdoel, et cetera). De resultaten bleken zelfs niet consistent over tests in hetzelfde vakgebied met soortgelijke kenmerken. Desalniettemin signaleren de auteurs de volgende algemene trends.
Van de 23 studies op het gebied van rekenen-wiskunde vonden er veertien geen verschil tussen PPT en CBT, bij acht was CBT moeilijker en bij slechts bij één studie was PPT moeilijker. Alhoewel de uitkomsten dus niet eensluidend waren, bleek de meerderheid van de toetsen voor rekenen-wiskunde vergelijkbaar over de twee afnamemodi.
In sommige van de studies die wel verschillen vonden, is gekeken naar itemtypen die het moduseffect zouden kunnen verklaren (Ito & Sykes, 2004; Sandene et al., 2005; Johnson & Green, 2006; Bennett et al., 2008; Keng, McClarty & Davis, 2008). Het gaat dan om items:

- die uit verschillende delen bestaan en waarbij de kandidaat moet scrollen om het hele item te zien;
- die mede tot doel hebben om te bepalen hoe effectief de kandidaat een fysiek hulpmiddel kan manipuleren, zoals een liniaal of een gradenboog;

- waarbij de leerling een tekening moet maken, een grote hoeveelheid tekst moet invoeren of een wiskundige formule moet produceren;
- waarbij de leerling grafisch materiaal of geometrische manipulaties moet produceren;
- die een uitgebreide handleiding (Engels: *tutorial*) of uitgebreide item-specifieke aanwijzingen voor het beantwoorden vereisen;
- die het gebruik van kladpapier vereisen (en waarbij de leerling de vraag dus niet 'uit het hoofd' kan beantwoorden):
- waarbij het itemformat ingrijpend gewijzigd moet worden om het item geschikt te maken voor presentatie via het beeldscherm;
- die veronderstellen dat de schermresolutie gelijk is op alle computers waarop de toets wordt afgenomen.

Verder komt uit de review naar voren dat leerlingen over het algemeen hogere scores op open vragen behalen wanneer de manier waarop de toets wordt afgenomen overeenkomt met de manier waarop de leerstof in de klas is onderwezen en getoetst.
Daarnaast blijken zelfs zeer kleine moduseffecten grote praktische consequenties te kunnen hebben. Een moduseffect van slechts één punt op een eindtoets in Texas kan ertoe leiden dat een groot aantal kandidaten zakt en dus geen diploma krijgt.

**Kingston (2009)**
In een meta-analyse analyseerde Kingston (2009) de resultaten van 81 vergelijkingsstudies met meerkeuzetoetsen, uitgevoerd gedurende de periode 1997 tot 2007 bij leerlingen in grade 1 tot en met 12. Voor het vakgebied wiskunde werden 31 studies geanalyseerd. PPT wiskunde werden iets beter gemaakt dan CBT wiskunde. De gemiddelde effectgrootte was met -.06 echter zeer klein (met een 95%-betrouwbaarheidsinterval van -.10 tot -.02). Kingston noemt het waarschijnlijk dat dit kleine verschil te maken heeft met het verschillend gebruik van kladpapier. Zie hiervoor ook paragraaf 3.4.

**Individuele studies met gegevens over vergelijkbaarheid van de rangordening**
Poggio, Glasnapp, Yang en Poggio (2005) legden leerlingen in grade 7 een PPT- en een CBT-versie van een meerkeuzetoets wiskunde voor. Zij vonden geen verschil in gemiddeld prestatieniveau en de gemiddelde correlatie tussen de scores op beide versies bedroeg .96.
De Oregon Department of Education (2007) vond geen moduseffect op de prestaties van wiskundetoetsen voor verschillende grades. De correlaties tussen de PPT- en CBT-versies varieerden over de grades van .70 tot .83.
Csapó, Molnár en Tóth (2009) vergeleken PPT- en CBT-versies van drie deeltests op het gebied van inductive reasoning. Op de onderdelen number analogies en number series scoorde PPT hoger dan CBT, maar op verbal analogies deed CBT het beter dan PPT. De correlaties tussen de beide testversies bedroeg .79 voor de totaalscores over alle drie de onderdelen, .62 en .42 voor respectievelijk number analogies en number series en .80 voor verbal analogies.

## 3.2 Het toekennen van partial credit aan antwoorden

**Inleiding**

In een papieren examen rekenen-wiskunde is het gebruikelijk dat kandidaten beloond worden voor gedeeltelijk goede antwoorden. De beoordelaar kent het volledige aantal punten toe als het antwoord helemaal goed is en kent deelscores (Engels: *partial credit*) toe als het antwoord gedeeltelijk goed is (McGuire & Youngson, 2002). Hoewel er voor doeleinden van formatieve toetsing in oefenomgevingen subtielere scoringswijzen beschikbaar zijn (o.a. DWO, 2013), is geautomatiseerde toekenning van *partial credit* aan gedeeltelijk goede antwoorden op open vragen in de meeste ons bekende digitale high-stakes wiskunde-examens niet gebruikelijk. Vrijwel altijd wordt alleen het eindantwoord automatisch gescoord als goed (één punt) of fout (nul punten). Bij het vakgebied rekenen-wiskunde is de meest voor de hand liggende vraag: “In hoeverre is het mogelijk de antwoorden van kandidaten in een automatisch gescoord CBT-examen op dezelfde manier te beoordelen als menselijke beoordelaars bij een papieren examen zouden doen?” (McGuire & Youngson, 2002; Ashton et al., 2006).

Examens rekenen-wiskunde waarbij de vergelijkbaarheid van PPT- en CBT-versies tot op heden het best gerealiseerd is, zijn CBT-examens met alleen meerkeuzevragen (Bennett, 2003). Dat is niet zo verwonderlijk, aangezien beide afnamemodi van het meerkeuze-examen doorgaans dichotoom gescoord worden. Echter, examens rekenen-wiskunde bevatten meestal meer open vragen dan meerkeuzevragen. Een antwoord op een open vraag kan vele vormen aannemen, zoals een numerieke score, een formule of een grafiek. Het antwoord op een wiskundevraag vereist doorgaans meerdere denkstappen of oplossingsprocessen waarvan de kandidaat er één of meer goed uitgevoerd kan hebben. Hoewel er uitzonderingen zijn (zie hierna) is het zonder speciale aanpassingen in een CBT-examen niet mogelijk om kandidaten voor gedeeltelijk goede antwoorden punten toe te kennen: de computer kent het antwoord alleen het maximale aantal punten toe als het antwoord helemaal goed is en als dat niet het geval is krijgt de kandidaat nul punten. De CBT-versie van het examen wijkt dan af van de PPT-versie waarbij menselijke beoordelaars deelscores toekennen aan gedeeltelijk goede antwoorden.

Docenten en kandidaten beschouwen de geautomatiseerde alles-of-niets scoring van antwoorden op open vragen in een CBT vaak als onrechtvaardig. De score nul wil immers niet zeggen dat de kandidaat geen enkele vaardigheid bezit of geen enkel leerdoel bereikt heeft. In een CBT kan een reken- of slordigheidsfoutje aan het begin van het oplossingsproces er bijvoorbeeld toe leiden dat de kandidaat voor het hele antwoord nul punten krijgt, ook al heeft de kandidaat de vereiste strategie of formule volledig correct toegepast. En als aanwezige partiële kennis en vaardigheden niet herkend en beloond worden, kunnen vraagtekens worden geplaatst bij de validiteit van de meting van wiskundige vaardigheden. Het wekt dan ook geen verbazing dat wiskundedocenten het toekennen van deelscores aan gedeeltelijk goede (tussen)antwoorden prefereren boven een alles-of-niets scoring van alleen het eindantwoord. Veel docenten vragen zich af in hoeverre kandidaten hogere scores behaald zouden hebben als deelscores wel beloond zouden zijn en hoeveel punten de kandidaten dan hoger zouden scoren; bovendien vragen zij zich af of er een manier is om kandidaten te compenseren voor de ten onrechte niet gekregen deelscores (zonder punten weg te geven) (Darrah et al., 2010). Bij kandidaten kan de geautomatiseerde alles-of-niets scoring van alleen het eindantwoord tot bezorgdheid, wantrouwen en frustraties leiden, onder meer vanwege het

onpersoonlijke en anonieme karakter van de volledig geautomatiseerde beoordeling. Kandidaten kunnen zich benadeeld voelen omdat zij het idee hebben punten mis te lopen die zij wel gekregen zouden hebben als hun antwoorden en uitwerkingen op papier door een menselijke beoordelaar nagekeken zouden zijn (McGuire & Johnson, 2002; Darrah et al., 2010).

De vraag naar effect van het al dan niet toekennen van *partial credit* aan gedeeltelijk goede prestaties is mogelijk ook van belang voor de interpretatie van de resultaten op de huidige Rekentoets VO. De tegenvallende resultaten zoals geconstateerd door de Commissie Bosker (2014) zijn mogelijk te verklaren vanuit het gegeven dat gedeeltelijk goede antwoord in de huidige digitale examinering geen punten opleveren. Onder constant houding van de maximumscore (en het antwoordmodel) zou *partial credit* scoring logischerwijs tot hogere scores moeten leiden dan alles-of-niets scoring van alleen het eindantwoord (uiteraard onder de aanname dat gedeeltelijk goede antwoorden inderdaad voorkomen). Dat het percentage goed op het examen door het toekennen van deelscores hoger wordt, betekent natuurlijk niet dat de kandidaten dan vaardiger zijn geworden. Het betekent wel dat de resultaten op het examen dan minder snel als tegenvallend geïnterpreteerd zullen worden. Daarnaast zijn er wellicht implicaties voor de referentiecesuren en de manier waarop deze via standaardsetting bepaald worden. In de begeleidende brief aan de minister constateert de Commissie Bosker (2014) dat de referentiecesuren voor de niveaus 2F en 3F thans nog te hoog liggen en dus bijstelling behoeven. Een mogelijke reden waarom de lat nu zo hoog ligt, is het gegeven dat de referentiecesuren bepaald zijn zonder gedeeltelijk goede antwoorden in de standaardsetting te verdisconteren.

In deze paragraaf gaan we na wat er bekend is over de volgende twee onderzoeksvragen:

- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met het al dan niet het toekennen van *partial credit* aan gedeeltelijk goede antwoorden?
- In hoeverre is het mogelijk de resultaten van de leerlingen op een CBT-examen op eenzelfde manier te beoordelen als de gebruikelijke *partial credit* beoordeling in een PPT-examen?

**Resultaten: reviews en meta-analyses**

Er is nog nauwelijks onderzoek gedaan naar de mogelijkheden van geautomatiseerde scoring van antwoorden op wiskundige problemen waarbij kandidaten beloond worden voor gedeeltelijk goede antwoorden (Darrah et al., 2010). We hebben geen review of meta-analyse gevonden waarin resultaten van ontwikkelingsgericht onderzoek op een handzame wijze gesynthetiseerd worden. Noodgedwongen volstaan we met een verhalende beschrijving van individuele studies. We beperken ons in hoofdzaak tot het beschrijven van de belangrijkste onderzoeksresultaten.

**Resultaten: individuele studies**

**Beevers, McGuire, Stirling en Wild (1995)**
In een van de eerste experimenten met geautomatiseerde *partial credit* scoring in een CBT wiskunde probeerden Beevers et al. (1995) een oplossing te vinden voor een bekend nadeel van CBT met open vragen: het probleem van het niet kunnen waarderen van gedeeltelijk goede antwoorden. Om *partial credit* beoordeling mogelijk te maken, ontwierpen de onderzoekers een CBT-versie waarbij elke toetsvraag werd opgedeeld in twee à vier deelvragen. De leerlingen konden per toetsvraag kiezen of

zij van *Steps* gebruik wilden maken door de knop *More steps* al dan niet aan te klikken. Om de maximumscore te behalen, moest de leerling alle deelvragen goed beantwoord hebben. De keuze voor *Steps* leidde dus niet automatisch tot aftrek van punten. De keuze voor *Steps* bleef echter niet geheel ongestraft, aangezien de toets een tijdslimiet kende en de stapsgewijze bevraging meer tijd bleek te kosten dan de reguliere bevraging. Zoals verwacht bleken goede leerlingen de toetsvragen vaak snel te beantwoorden zonder de tussenstappen te gebruiken. Sommigen van hen liepen daardoor echter punten mis die zij wel gekregen zouden hebben als een menselijke beoordelaar hun antwoorden op papier beoordeeld zou hebben. Zwakkere leerlingen kozen veelvuldig voor *Steps*, bijvoorbeeld als zij niet onmiddellijk in staat waren een 'beginnetje' te vinden of tussentijds 'vastliepen' bij het vinden van een antwoord. Zij verdienden daardoor punten die beoordelaars van een PPT ook gegeven zouden hebben. Het kwam maar weinig voor dat de zwakkere leerlingen niet voor *Steps* kozen terwijl zij dat beter wel hadden kunnen doen.
Deze keuze om *Steps* al dan niet te gebruiken vereist toetsvaardigheden die bij een PPT niet nodig zijn. De kandidaat moet het voordeel van het uitzicht op *partial credit* namelijk afwegen tegen het nadeel van een toename van de afnametijd. De onderzoekers doen dan ook de aanbeveling om deze nieuwe toetsvaardigheden terdege te trainen.

**McGuire, Youngson, Korabinski en McMillan (2002)**
Aanleiding van het onderzoek van McGuire, Youngson, Korabinski en McMillan (2002) waren de problemen met de alles-of-niets scoring van alleen het eindantwoord in traditionele CBT. Als het scoringsalgoritme de kandidaat nul punten toekent, betekent dat niet automatisch dat hij of zij over geen enkele wiskundige vaardigheid beschikt.
Modernere CBT's die mogelijkheid bieden tot stapsgewijze aanbieding van vragen, kunnen partial credit toekennen, bijvoorbeeld voor het vinden van het juiste begin van het antwoord (Engels: *the correct start to the answer*), voor de correcte toepassing van de formule (ook al gebruikte de kandidaat in de formule een verkeerd getal vanwege een reken- of slordigheidsfoutje eerder in de opgave) of voor stapelfouten (Engels: *follow-through errors*) waarmee wordt voorkomen dat dezelfde fout de kandidaat meer dan één keer aangerekend wordt.

In een zeer invloedrijk onderzoek vergeleken McGuire, Youngson, Korabinski en McMillan (2002) de digitale beoordeling van wiskundige vaardigheden met en zonder recht te doen aan gedeeltelijk goede antwoorden. In een experimenteel onderzoeksontwerp vergeleken zij de volgende drie versies van dezelfde wiskundetoets die verschilden in de hoeveel hulp die de kandidaat geboden werd:

- *No Steps* format (NS): een traditionele CBT-versie met geautomatiseerde dichotome scoring van alleen het eindantwoord. In het *No Steps* format werden gedeeltelijk goede antwoorden dus niet beloond.
- *Compulsory Steps* (CS) format: een CBT-versie met zogenoemde *Steps* waarbij elke examenvraag in een aantal deelvragen was opgedeeld. Om de maximumscore te behalen, moest de kandidaat zowel de antwoorden op de deelvragen als het antwoord op de eindvraag goed hebben. De deelvragen kwamen bij benadering overeen met de "*method the student would have to go through in order to solve the question*" (p. 224). Omdat de kandidaat niet voor *Steps* kon kiezen, werd deze versie aangeduid met de term *Compulsory Steps.*
- *Optional Steps* (OS) format: een hybride CBT-versie die het midden hield tussen NS en CS. De kandidaat kreeg de examenvraag eerst in het NS-format te zien en kon daarna desgewenst kiezen voor Steps. Had de kandidaat eenmaal voor Steps gekozen, dan moesten alle deelvragen

correct beantwoord worden om de maximumscore te behalen. Net als in het hiervoor besproken onderzoek van Beevers, McGuire, Stirling en Wild (1995) leidde de keuze voor Steps niet automatisch tot aftrek van punten. De keuze voor Steps bleef echter niet geheel ongestraft, aangezien de toets een tijdslimiet kende, de afnametijd daar *Steps* toenam en de kandidaat met tijdnood te maken kon krijgen.

In het OS-format werd 30% van de vragen geprobeerd zonder van Steps gebruik te maken. Van deze vragen werd 42% goed beantwoord en 58% fout. Van de kandidaten gebruikte 28% Steps bij alle vragen, 54% gebruikte Steps bij ten minste één vraag en 8% gebruikte Steps helemaal niet.

De antwoorden op de vragen van de drie CBT-versies werden automatisch met de computer gescoord. Als de computer een goed antwoord niet herkende, werden de scores handmatig hersteld. Om de praktijk van het toekennen van deelscores bij een gewone PPT zo goed mogelijk na te bootsen, werden de uitwerkingen op kladpapier door menselijke beoordelaars beoordeeld. Dit resulteerde in de volgende zeven 'condities' die met elkaar vergeleken werden:

- *NS: No Steps* waarbij alleen het eindantwoord automatisch dichotoom gescoord werd;
- *NSW: No Steps* waarbij een menselijke beoordelaar de uitwerkingen op kladpapier volgens de *partial credit* methode beoordeelde;
- *CS: Compulsory Steps* automatisch met *partial credit* gescoord;
- *CSPC: Compulsory Steps* waarbij een menselijke beoordelaar de uitwerkingen op kladpapier volgens de *partial credit* methode beoordeelde;
- *OS: Optional Steps* automatisch met *partial credit* gescoord;
- *OSPC: Optional Steps* waarbij een menselijke beoordelaar ALLEEN DE VRAGEN waarbij de kandidaten ervoor kozen geen gebruik te maken van Steps volgens de *partial credit* methode beoordeelt (op basis van de uitwerkingen op kladpapier);
- *OSPC+S: Optional Steps* waarbij een menselijke beoordelaar ALLE VRAGEN volgens de *partial credit* methode beoordeelde (op basis van de uitwerkingen op kladpapier).

Het onderscheid tussen CS en SCPC en dat tussen OSPC en OSPC+S motiveren McGuire et al. (2002) als volgt: "*The main reason for a difference between CS and CSPC marks for a particular candidate was that if they gave a wrong answer to, let us say, the first part of the question and then subsequently used the right method to the remaining parts, then the computer would give no marks for the remaining parts whereas partial credit would normally accrue in paper-based examinations. The same main reason applied to a difference between OS and OSPC marks for any particular candidate. The differences between NS and NSW marks (and OSPC and OSPC+S marks) could not be assigned in such a simple way as the partial credit was awarded for making variable amounts of progress through each question*." (p. 226).

De via NSW verkregen scores werden geacht de menselijke beoordeling van antwoorden op papier het beste te benaderen. De NSW-beoordeling met *partial credit* bleek tot aanzienlijk hogere scores te leiden dan NS (dit wil zeggen: de gebruikelijke dichotome scoring van alleen het eindantwoord). McGuire et al. (2002) trekken hieruit de conclusie dat "… *there is absolutely no doubt that the basic NS format is not a suitable alternative to paper-based examinations*" (p. 226).
De onderzoekers vergeleken ook NSW met OS en CS. De OS-scores waren gemiddeld iets lager dan de NSW-scores, maar dit verschil was statistisch gezien niet significant (op 10%-niveau). Ook de CS-

scores waren iets hoger dan de NSW-scores, maar ook dit verschil was statistisch gezien niet van betekenis. Kennelijk geven de optionele en verplichte Steps-methodes een goede benadering van de traditionele PPT-beoordeling met toekenning van partial credit door menselijke beoordelaars. Het lijkt dus goed mogelijk om de 'gewone' arbeidsintensieve PPT-beoordeling met deelscores in een automatisch gescoorde CBT met *Steps* na te bootsen

Het onderzoek bood ook uitsluitsel over eventuele verschillen in de gemiddelde scores zoals verkregen met de optionele en verplichte Steps-methode. Er waren drie vergelijkingen mogelijk. De vergelijking CS versus OS gaf significantie te zien waarbij CS gemiddeld tot iets hogere scores leidde dan OS. De vergelijking CSPC versus OSPC was eveneens significant waarbij CSPC wederom tot hogere scores leidde dan OSPC. Hieruit concluderen de onderzoekers dat het erop lijkt dat de verplichte steps-methode de kandidaten hulp biedt bij de keuze van de juiste strategie (die zij bij een gewone PPT niet zouden hebben gekregen). De vergelijking CSPC versus OSPC+S leverde daarentegen geen significantie op. Dit betekent dat het toekennen van partial credit bij alle vragen de CS-scores weer op gelijke hoogte brengt met de OS-scores.

Tot slot werd NS vergeleken met OS en CS. Hier waren twee vergelijkingen mogelijk. De vergelijking NS versus OS gaf significantie te zien waarbij NS tot lagere scores leidde dan OS. Ook de vergelijking NS versus CS leverde significantie op waarbij NS lager scoorde dan OS. Dit laat volgens de onderzoekers zien dat het gebruik van de optionele en verplichte Steps-methode de kandidaten meer mogelijkheden biedt om hun kennis ten toon te spreiden dan de gebruikelijke CBT-toetsing waarbij alleen het eindantwoord dichotoom gescoord wordt.
McGuire et al. (2002) zijn er zich overigens van bewust dat de Steps-methoden een deel van het antwoord kunnen weggeven. Daardoor zouden examenvragen met Steps een beroep kunnen doen op andere leerdoelen dan de traditionele papieren beoordeling aan de hand van een antwoordmodel. Zij pleiten daarom voor vervolgonderzoek naar het toekennen met aftrekpunten als straf voor het gebruik van Steps. Wel zouden kandidaten dan gewaarschuwd moeten worden dat de keuze voor Steps niet alleen tijdverlies kan betekenen, maar ook aftrekpunten.

**Fiddes, Korabinski, McGuire, Youngson en McMillan (2002)**
In een onderzoek dat parallel liep aan dat van McGuire, Youngson, Korabinski en McMillan (2002) vergeleken Fiddes, Korabinski, McGuire, Youngson en McMillan (2002) drie versies van dezelfde wiskundetoets:

- een PPT-versie, beoordeeld volgens de *partial credit* methode;
- een CBT-versie, beoordeeld volgens de alles-of-niets methode van alleen het eindantwoord;
- een zogenoemde *Reversed Translation* versie (RT), dit wil zeggen een screendump van de CBT-versie (ofwel een vanuit de CBT-versie 'terugvertaalde' PPT-versie).

In de RT-versie was bij elke toetsvraag ruimte gereserveerd voor kladwerk, zodat de leerlingen hun berekeningen en dergelijk konden uitvoeren op een manier die vergelijkbaar was met de PPT-versie. Om het effect van herformulering onderzoeksmatig te kunnen onderscheiden van dat van de afnamemodus, werd de RT-versie op twee manieren beoordeeld:

- *RTC*. Bij de zogenoemde *RTC-marking* werd alleen beoordeeld of het eindantwoord goed of fout was, net als bij de CBT-versie gebeurd was;

- *RTW*. Bij de zogenoemde *RTW-marking* vond plaats volgens de *partial credit* methode waarbij de uitwerkingen op kladpapier in de beoordeling betrokken werden en de leerling punten kon behalen voor gedeeltelijk goede antwoorden, zoals ook bij de PPT het geval was.

Het onderzoekontwerp stelde de onderzoekers in staat om het effect van de afnamemodus te scheiden van het effect van het herformuleren van de examenvragen.

**Effect afnamemodus**

Omdat de formulering van de vraagstelling en de plek om het antwoord te noteren in de CBT- en RT-versie exact gelijk waren en beide versies op precies dezelfde wijze beoordeeld werden via de goed-of-fout methode, geeft de vergelijking CBT-RTC inzicht in het mogelijke effect van de afnamemodus. De afnamemodus bleek geen noemenswaardig effect op de prestaties te hebben. Kennelijk maakt het voor de prestaties niet uit of de leerlingen de originele CBT maken of een papieren screendump van deze versie zolang hun antwoorden maar volgens de alles-of-niets benadering gescoord of beoordeeld worden.

**Effect herformulering**

Voordat de vragen van een PPT geschikt zijn voor gebruik in een CBT, moeten ze vaak geherformuleerd en in een andere lay-out gezet worden. Fiddes et al. (2002) noemen dit het herformuleringseffect. De vergelijking van de PPT- met de RTW-versie, die beide volgens de partial credit methode beoordeeld werden, geeft inzicht in het eventuele effect van het herformuleren van de toetsvragen. De herformulering van de toetsvragen bleek een middelgroot effect op de toetsresultaten te hebben gehad (effectgrootte .48), waarbij de RTW-versie beter gemaakt werd dan de PPT-versie. De onderzoekers schrijven dit verschil toe aan de herformulering van de vragen en de verschillen in het aantal opgaven per pagina op papier en het beeldscherm.
Een alternatieve verklaring verwijst naar het gegeven dat de leerlingen hun uitwerkingen op kladpapier in het PPT-format onder elkaar opschreven, zodat de beoordelaar gemakkelijk kon zien waar de fouten zich voordeden. In het RTW-format konden de leerlingen hun antwoorden daarentegen niet op een lineaire manier opschrijven. Als de leerling bijvoorbeeld in het RTW-format twee antwoorden gaf waarvan er één fout en één goed was, was veel minder duidelijk welk antwoord de leerling als het goede antwoord bedoeld had. Bijgevolg gaven de beoordelaars de leerlingen in het RTW-format vaker het voordeel van de twijfel dan in het PPT-format. De onderzoekers merken op dat deze kwestie nader onderzoek vereist. In het onderzoek werd geen gebruik gemaakt van stapsgewijze toetsing die de onderzoekers aanbevelen als een geschikte manier om gedeeltelijk goede antwoorden in een CBT te belonen.

**Ashton, Beevers, Korabinski, Youngson en Martin (2006)**

Ashton, Beevers, Korabinski, Youngson en Martin (2006) borduurden voort op het baanbrekende onderzoek van McGuire et al. (2002). Laatstgenoemden vonden geen verschil tussen de scores die behaald waren op een CBT met ingebouwde stapsgewijze bevraging en een gewone PPT waarbij menselijke beoordelaars partial credit toekenden aan gedeeltelijk goede antwoorden op papier. McGuire et al. (2002) waren zich er echter van bewust dat stapsgewijze bevraging zou kunnen leiden tot het weggeven van de juiste oplossingsstrategie. Dit werd als minder wenselijk ervaren, omdat het vinden van de juiste oplossingsstrategie een belangrijk leerdoel is van het vak wiskunde in het voortgezet onderwijs. Het onderzoek van Ashton, Beevers, Korabinski, Youngson en Martin (2006)

had tot doel een oplossing voor dit probleem te vinden. Zij onderzochten de volgende twee manieren om partial credit in een CBT toe te kennen:

- Het werken met aftrekpunten. Om tegemoet te komen aan het bezwaar dat de kandidaten die voor Steps kozen in sommige gevallen de strategie cadeau kregen, werden de punten die zij anders voor de strategie hadden gekregen op hun score in mindering gebracht; zij konden dus alleen de resterende punten behalen. Hiermee hoopten de onderzoekers het antwoordmodel van de originele PPT zo goed mogelijk na te bootsen (Ashton & Youngson, 2004).
- Het geven van onmiddellijke feedback op kleine foutjes. De tweede aanpak was gebaseerd op de observatie dat veel foute antwoorden in een traditionele CBT het gevolg waren van kleine reken- of slordigheidsfoutjes die in een PPT nauwelijks tot puntenaftrek geleid zouden hebben. Door de kandidaat op deze kleine foutjes te attenderen, zou deze de kans krijgen om de fout te traceren en het foute antwoord door het goede te vervangen. De feedback werd gegeven in de vorm van vinkjes en kruisjes (Engels: *ticks and crosses*) voor respectievelijk een goed en een fout antwoord. In de feedbackversie konden de kandidaten niet voor *Steps* kiezen. Zie voor functioneren van deze feedback de bespreking in paragraaf 3.5.

Ashton et al. (2006) ontwikkelden drie versies van hetzelfde wiskunde-examen (waarbij elke versie uit twee deeltoetsen bestond):

- *S-versie*: een CBT met optionele Steps waarbij de kandidaten desgewenst konden kiezen voor stapsgewijze bevraging met puntenaftrek als straf;
- *T-versie*: een CBT zonder Steps maar met onmiddellijke feedback via vinkjes en kruisjes op het scherm;
- *R-versie*: een screen dump van de CBT-versie zonder Steps**.**

Er werden twee trials uitgevoerd. In de eerste trial werden vragen uit het *Advanced Higher Mathematics* examen gebruikt. In de tweede trial gebruikte men vragen uit het *Higher Mathematics* examen dat qua niveau vergelijkbaar is met het A level examen in Engeland. In de eerste trial kregen de kandidaten alleen de S- en R-versies voorgelegd en in de tweede trial alle drie versies.

Alvorens de CBT-versie met optionele Steps te maken, werden de kandidaten eraan herinnerd dat de keuze voor Steps betekende dat zij niet de maximale score konden behalen. Kandidaten die de feedback-versie (T) maakten, werden geattendeerd op de vinkjes en kruisjes die zouden verschijnen nadat zij hun antwoord hadden ingevoerd. Allen werd gevraagd hun uitwerkingen op kladpapier te noteren.

Bij de afname van de CBT's bleken zich enkele uitvoeringsproblemen voor te doen. De beide CBT's bleken meer afnametijd te vergen dan de PPT. De kandidaten die een CBT-versie maakten, kwamen daardoor vaak in tijdnood en kregen het examen niet af binnen de afnametijd van één uur. De onderzoekers noemen de volgende oorzaken:

- Bij de CBT's kostte het opstarten van het examen meer tijd dan bij de PPT. De kandidaten moesten verschillende keren inloggen (in de computer, het Internet en het examen) en de weg vinden naar de juiste webpagina van het examen.
- Het intypen van de antwoorden is tijdrovender dan het noteren op papier.
- Kandidaten die bij de CBT met optionele Steps voor Steps kozen, gebruikten meer afnametijd dan degenen die daar niet voor kozen.

- In de feedback-versie onderbraken studenten wellicht hun werk om een fout antwoord te checken.

Een ander uitvoeringsprobleem had te maken met de invoer van de antwoorden op het scherm (zie ook de bespreking van de resultaten met de T-versie in paragraaf 3.5). De computer rekende soms antwoorden fout die een menselijke beoordelaar wel (gedeeltelijk) goed gerekend zou hebben. Een voorbeeld is het op zich juiste antwoord '*twenty one and a third*' waarbij de kandidaten niet altijd wisten hoe zij de samengestelde breuk (Engels: *fractional part*) moesten invoeren opdat de computer deze zou herkennen. Een ander voorbeeld is het vergeten van haakjes, bijvoorbeeld 'sin 2x' versus 'sin(2x)'. Omwille van de vergelijkbaarheid van de CBT- en PPT-versies werden invoerfouten die bij een PPT niet tot puntenaftrek zouden leiden handmatig hersteld.

*Advanced Higher Mathematics*
Bij de *Advanced Higher Mathematics* trial waren er twee deeltoetsen. Op de eerste deeltoets behaalden de kandidaten hogere scores op de S-versie (CBT met optionele Steps) dan op de R-versie (screendump van de CBT zonder Steps). Een vraag-bij-vraag analyse liet zien dat dit vooral veroorzaakt werd door één examenvraag die veel mogelijkheden bood tot het maken van rekenfoutjes die dan doorwerkten in het vervolg van de vraag (Engels: *follow-through errors*). In de papieren R-versie kregen de kandidaten voor deze vraag veel *follow-through partial credit*, terwijl de kandidaten in de S-versie daarvoor vrijwel altijd nul punten behaalden.

Op de tweede deeltoets was het verschil tussen de S- en R-versie niet significant. De analyse van de resultaten per vraag liet zien dat deze tweede deeltoets net als de eerste deeltoets ook een vraag bevatte met veel mogelijkheden tot het maken van stapelfouten. Anders dan bij de eerste deeltoets het geval was, verschilden de gemiddelden op de S- en R-versie van deze vraag niet van elkaar. Een nadere inspectie bracht van de antwoorden op deze vraag bracht aan het licht dat een foutje in het begin van de vraag het vrijwel onmogelijk maakte om de vraag verder te beantwoorden, waardoor de kandidaten voor het vervolg van de vraag weinig *partial credit* konden verdienen (zowel in de CBT met steps als de PPT). Dit in tegenstelling tot de eerste deeltoets waar kandidaten het geluk hadden de vraag verder te kunnen beantwoorden, al was dat dan wel met de verkeerde getallen.

De onderzoekers onderzochten ook het gebruik van Steps in de S-versie. De kandidaten bleken maar weinig gebruik te maken van Steps, mogelijk omdat zij zich realiseerden dat het gebruik ervan tot puntenaftrek zou leiden. Het waren vooral de zwakkere kandidaten die de meeste behoefte hadden aan hulp die dan voor Steps kozen.

Volgens de onderzoekers bevestigt het onderzoek de hypothese dat - met uitzondering van die ene vraag in de eerste deeltoets - beide versies tot dezelfde scores leiden en bovendien een beroep doen op soortgelijke kennis.

*Higher Mathematics*
In de *Higher Mathematics trial* bracht de vergelijking van de gemiddelde scores op de S-, T- en R-versie geen significante verschillen aan het licht. Bij een enkele vraag was de gemiddelde score op de S-versie wel lager dan die op de R-versie. Een nadere analyse liet zien dat de door Steps geboden hulp in de S-versie onvoldoende compenseerde voor de deelscores die de kandidaat bij de R-versie

zou hebben gekregen. Voor een bespreking van de onderzoeksresultaten met betrekking tot het functioneren van de T-versie wordt verwezen naar paragraaf 3.5.

Op basis van de onderzoeksresultaten brachten de onderzoekers ook een verbetering aan in de S-versie (CBT met steps). In de trials wisten kandidaten al wel dat zij punten verliezen als zij voor Steps kiezen, maar nog niet hoeveel dat er zouden zijn. In de verbeterde versie krijgen kandidaten alvorens eventueel voor Steps te kiezen feedback over het maximale aantal resterende punten dat zij dan nog kunnen behalen. Voor gedetailleerde informatie over de verbeteringen die in de CBT met steps zijn aangebracht wordt verwezen naar de *Good Practice Guide in Question and Test Design* (Pass-it, 2002). De resultaten van het onderzoek maken, aldus de onderzoekers, aannemelijk dat een verbeterde CBT-versie met steps de normale papieren versie van het wiskunde-examen kan vervangen. Een vervolgonderzoek onder high-stakes condities zou dit moeten bevestigen.

**Darrah, Fuller en Miller (2010)**

Darrah, Fuller en Miller (2010) ontwierpen een CBT voor het examineren van wiskundige vaardigheden bij een *Introductory Calculus Course*. Het digitale examen bood zoals te doen gebruikelijk geen mogelijkheden tot het honoreren van gedeeltelijk goede antwoorden. Anders dan bijvoorbeeld Ashton et al. (2006) deden Darrah et al. (2010) geen pogingen om de deelscorebenadering in hun CBT in te bouwen. In plaats daarvan legden zij de kandidaten aan het einde van het semester een beknopte aanvullende toets voor over de leerstof van het hele semester. Anders dan bij de CBT werden gedeeltelijk goede antwoorden bij deze zogenoemde *Super Quiz* wel beloond. Het doel van deze eindtoets was de leerlingen in de gelegenheid te stellen de bij de CBT gemiste deelscores alsnog te behalen. Aan het begin van het CBT-examen werd de kandidaten verteld dat zij al hun uitwerkingen en dergelijke op kladpapier moesten noteren (en inleveren voordat zij de examenzaal verlieten). Op deze manier kon hun werk worden beoordeeld als ware het een PPT waarbij de antwoorden met de hand werden nagekeken volgens de *partial credit* methode. De veronderstelling was dat de op de *Super Quiz* behaalde extra punten vergelijkbaar zouden zijn met de extra punten die de beoordeling van het kladwerk volgens de *partial credit* methode zou opleveren.

De onderzoekers vergeleken de geautomatiseerde alles-of-niets scoring van alleen het eindantwoord op de CBT met de *partial credit* beoordeling van de uitwerkingen op kladpapier. Zoals verwacht, bleek de *partial credit* beoordeling tot aanmerkelijk hogere scores te leiden dan de dichotome CBT-scoring. Het verschil tussen de gemiddelden voor de vier deeltoetsen varieerde van 3.21 tot 6.10 scorepunten (bij een gemiddelde toetsscore variërend van 67.76 tot 80.35 en een standaarddeviatie van de *partial credit* beoordeling van 9.78; de standaarddeviatie van de CBT vermelden de onderzoekers helaas niet).

Daarnaast vergeleken de onderzoekers de scores op de *Super Quiz* met de *partial credit* beoordeling van de uitwerkingen op kladpapier. Zij vonden een significant verschil in het voordeel van de *Super Quiz*. Zij concludeerden daaruit dat deze eindtoets te sterk compenseert voor de gemiste deelscores op de dichotoom gescoorde CBT.

Verder vroegen de onderzoekers zich af in hoeverre vaardige kandidaten evenveel profiteren van de *Super Quiz* en de *partial credit* beoordeling als de minder vaardige kandidaten. Daartoe verdeelden zij de kandidaten in twee groepen: de 20% beste en de 80% zwakste studenten. De groep beste

kandidaten bleek bij de *Super Quiz* veel meer van de mogelijkheid tot het behalen van bonuspunten te profiteren dan bij de *partial credit* beoordeling het geval was. De onderzoekers verklaren dit door te ernaar te verwijzen dat de *partial credit* beoordeling voor hen relatief weinig bonuspunten kon opleveren vanwege hun reeds relatief hoge scores op de CBT. De 20% beste studenten verzilverden tussen de 0% en 50% van de extra punten die de *partial credit* beoordeling maximaal kon opleveren. Omdat zij in de CBT meer vragen gemist hadden, konden de 80% zwakste kandidaten meer punten 'terugverdienen' dan de 20% beste kandidaten. Zij bleken dat echter niet te doen, noch bij de *Super Quiz*, noch bij de *partial credit* beoordeling van hun uitwerkingen op kladpapier. De onderzoekers concluderen hieruit dat de betere kandidaten meer profiteren van de *partial credit* beoordeling dan de zwakkere kandidaten, maar dat de rangordening van de kandidaten in de meeste gevallen gelijk was gebleven. Dit laatste bleek ook uit een vergelijking van de frequentieverdelingen van de cijfers op basis van de CBT en de *partial credit* beoordeling.

In het onderzoek is de studenten na afloop van het onderzoek gevraagd naar hun voorkeur voor papieren en digitale examinering. Op de vraag 'Als je zou kunnen kiezen, zou je dan de voorkeur geven aan een papieren examen boven een gecomputeriseerd examen' antwoordde 48% van de studenten met Ja, 25% met Nee en 27% had geen voorkeur. Van de studenten die de open vraag *Please make other comments about computerized exams* beantwoordden, noemde de helft het nadeel dat het digitale examen geen *partial credit* toeliet.

## 3.3 Mogelijkheden van item review

**Inleiding**

Volgens de bekende APA-richtlijnen (1986) dienen kandidaten de gelegenheid te krijgen om hun antwoorden te herzien. Wise en Plake (1989) verstaan onder item review a) het overslaan van vragen om deze later te beantwoorden, b) het bekijken van eerder gegeven antwoorden en c) het veranderen van eerder gegeven antwoorden. Verondersteld wordt dat item review een positieve bijdrage levert aan de validiteit van de meting (Vispoel, 2000; Schwarz, McMorris & DeMers, 1991). De kandidaat kan immers fouten corrigeren die het gevolg zijn van a) schrijf-, notatie-, type- en invoerfouten, b) het in eerste instantie verkeerd interpreteren van examenvragen, c) tijdelijke fluctuaties in het concentratievermogen, en c) het opnieuw overdenken van eerder gegeven antwoorden. Daarnaast kan het toestaan van item review toetsangst verminderen, met name bij laag vaardige leerlingen. Anders dan PPT biedt de eerste-generatie CBT geen mogelijkheden tot item review. Hierna bespreken we het gevonden onderzoek naar item review en het effect ervan op de prestaties. Hierbij merken we op dat de rekentoetsen anno 2015 ruime mogelijkheden bieden tot item review.

**Resultaten: reviews en meta-analyses**

**Mueller en Wasser (1977)**

Mueller en Wasser (1977) analyseerden de resultaten van achttien studies naar het effect van item review op de prestaties van papieren meerkeuzetoetsen. Zij rapporteren winst-verlies ratio's tussen de 2.3 : 1 en 5.3 : 1. Dit betekent dat veranderingen van een fout naar een goed antwoord twee à vijf keer zo vaak voorkomen dan veranderingen van een goed naar een fout antwoord. De conclusie was

dat item review een positieve bijdrage aan de prestaties levert en dat de meer vaardige leerlingen daar meer van profiteerden dan de minder vaardige leerlingen.

**Benjamin, Cavell en Shallenberger (1984)**
Benjamin, Cavell and Shallenberger (1984) analyseerden 33 studies naar item review bij papieren toetsen. Hun conclusie luidde als volgt: "*…after more than a half century of research on this topic*" *the evidence uniformly indicates that a) only a small percentage of answers are actually changed, b) the majority of answers are changed from wrong to right, c) most test takers are answer changers, and d) most answer changers are point gainers"* (p. 133).

**Resultaten: individuele studies**

**Lee en Hopkins (1985)**
Lee en Hopkins (1985) vergeleken een PPT- en CBT-versie van een redeneertoets wiskunde. De CBT bood geen mogelijkheden tot item review. De studenten behaalden hogere scores op de PPT dan de CBT. De onderzoekers concludeerden dat item review van invloed is op de prestaties en betoogden dat "*only software that allows the conveniences of paper-and-pencil tests, e.g., the ability to change answers and the ability to review past items, be used in future applications*" (p. 9).
Terzijde zij opgemerkt dat Russell, Goldberg en O'Connor (2003) melding maken van twee andere vroege studies - naast die van Lee en Hopkins uit 1985 - naar wiskundig redeneren. De resultaten ervan waren wisselend: afname per computer bleek zowel van positieve invloed (Johnson & Mihal, 1973) als van negatieve invloed (Lee, Moreno & Sympson, 1986) op de wiskundeprestaties. Zoals we hierna zullen zien, zijn de resultaten van Lee en Hopkins (1985) in diverse studies bevestigt (Vispoel, 1998; Wise & Plake, 1989; 2000; Vispoel, Wang, de la Torre, Bleiler & Dings, 1992).

**Vispoel (1988, 2000) en Vispoel, Wang, De la Torre, Bleiler en Dings (1992)**
Vispoel (1988, 2000) en Vispoel, Wang, De la Torre, Bleiler en Dings (1992) bestudeerden de antwoorden van kandidaten op een CBT-woordenschattoets voorafgaand en nadat zij de mogelijkheid tot item review hadden gekregen. De resultaten laten zien dat veel kandidaten hun antwoorden herzien; zij doen dat echter slechts bij een klein deel van de items. Vispoel (1998) vond dat 67% van de kandidaten een of meer antwoorden hadden gewijzigd. In het onderzoek van Vispool (2000) maakte bijna de helft van de kandidaten (45%) gebruik van de mogelijkheid tot item review, maar minder dan 4% bracht daadwerkelijk één of meer veranderingen aan.
Item review in een woordenschattoets heeft een positief effect op de prestaties. Vispoel (2000) vond bijvoorbeeld dat de kandidaten het antwoord ruim twee keer vaker van fout naar goed corrigeerden dan van goed naar fout (2.25 : 1). Wel had item review tot gevolg dat de afnametijd aanzienlijk toenam, al ging dit minder op voor hoog dan voor laag vaardige leerlingen. Naarmate kandidaten vaardiger waren, nam het aantal veranderde antwoorden af een steeg de fout-naar-goed ratio (4 : 1). Tot slot bleken de kandidaten de mogelijkheid tot item review zeer op prijs te stellen. Ook Vispoel et al. (1992) concludeerden dat het kunnen terugzien en veranderen van de antwoorden op een woordenschattoets een positief effect had op de prestaties. Daarnaast zorgde item review voor een kleine afname van de meetnauwkeurigheid en een toename van de totale afnametijd. En wederom hadden de kandidaten een sterke voorkeur voor item review.

**Wise en Plake (1989)**

Wise en Plake (1989) gingen na in hoeverre het ontbreken van de mogelijkheden tot item review in de eerste-generatie CBT de prestaties nadelig beïnvloedt. Als mogelijke oorzaken voor het onderstelde negatieve effect op de prestaties noemen zij: a) het niet kunnen overslaan van vragen om deze later te beantwoorden, b) het niet kunnen bekijken van eerder gegeven antwoorden en c) het niet kunnen veranderen van eerder gegeven antwoorden. Zij verwachtten dat meer flexibiliteit zou leiden tot hogere prestaties, maar ook tot een toename van de toetstijd. Het enige onderzoek dat Wise en Plake op dit terrein konden vinden, was het niet gepubliceerde promotie-onderzoek van Harvey (1989). Zij ontwikkelde een CBT waarin de drie hiervoor genoemde kenmerken geïmplementeerd waren en vergeleek deze versie met een CBT zonder deze kenmerken. Er werden geen significante verschillen in de prestaties op beide versie gevonden. De resultaten moeten echter met de nodige voorzichtigheid geïnterpreteerd worden omdat de motivatie van de studenten te wensen overliet.

**Schwartz, McMorris en DeMers (1991)**

Schwartz, McMorris and DeMers (1991) bestudeerden de redenen van kandidaten om antwoorden op toetsvragen te veranderen. De meerderheid van de leerlingen bleek dat te doen om wat de onderzoekers legitieme redenen noemen. Bijna de helft (45%) van de leerlingen zegt antwoorden te veranderen omdat zij de vraagstelling na herlezen beter begrepen, bijna een derde (31%) omdat zij de vraag daardoor beter konden overdenken en herconceptualiseren en een vijfde (20%) omdat zij zich daardoor meer informatie konden herinneren. Verder bleken de middelmatige en betere leerlingen meer te profiteren van item review dan de zwakkere leerlingen. Volgens de auteurs is het niet meer dan billijk dat leerlingen de kans krijgen hun werkelijke kennis te demonstreren door hun werk te controleren op invoerfouten en ontdekte fouten te verbeteren.

**Lunz, Bergstrom en Wright (1992)**

Lunz, Bergstrom en Wright (1992) ontwikkelden een CAT-examen medische technologie. Volgens een experimenteel onderzoeksontwerp wezen zij kandidaten random toe aan vier condities die van elkaar verschilden in de mogelijkheden van item review:

- *Skip*. In de zogenoemde *skip condition* kon de kandidaat bij elk toegewezen item ervoor kiezen de vraag al dan niet te beantwoorden. Verkoos de kandidaat een item niet te beantwoorden, dan kreeg hij of zij een andere item van vergelijkbare moeilijkheidsgraad over hetzelfde inhoudelijk gebied toegewezen. Eenmaal overgeslagen items mochten niet alsnog geprobeerd worden. Deze conditie bood de kandidaat maximale controle over het examen.
- *Review*. In de zogenoemde *review condition* maakte de kandidaat alle items, maar na het laatste item mochten zij hun antwoorden bekijken en zo nodig herzien. De onderzoekers merken op dat kandidaten menen recht te hebben op item review, en daar vaak in getraind zijn.
- *Defer*. Ook in de zogenoemde *defer condition* moest de kandidaat alle toegewezen items beantwoorden, maar zij konden het beantwoorden van een item uitstellen tot aan het einde van het examen. Was het laatste item gemaakt, kreeg de kandidaat de uitgestelde items alsnog aangeboden. De reeds beantwoorde items kreeg de kandidaat niet nogmaals te zien. De kandidaat kon dus zelf bepalen wanneer hij of zij het item wilde beantwoorden, maar beantwoordde wel alle items.
- *Non*. In de zogenoemde *non condition* had de kandidaat geen enkele controle over het examen. Elk item werd beantwoord op het moment van toewijzing, en de kandidaat mocht geen items

overslaan, uitstellen of terugkeren naar eerder aangeboden items. Deze vierde conditie biedt optimale psychometrische controle over het examen.

Hoe minder de kandidaat invloed op het examen kon uitoefenen, hoe lager de prestaties. De verschillen in gemiddelde prestaties tussen de vier condities waren echter slechts in één geval significant (te weten: de vergelijking van de skip- versus de non-condition).

In de *skip*-conditie maakte bijna twee derde (64%) van de kandidaten gebruik van de mogelijkheid om items over te slaan. Daarbij werd 9% van de items daadwerkelijk overgeslagen (d.w.z. geweigerd waarna de kandidaat een vervangend item kreeg aangeboden). Kandidaten in de *skip*-conditie presteerden significant hoger dan degenen die geen controle hadden over hun adaptieve toets. Ter verklaring verwijzen de onderzoekers naar het gevoel van veiligheid en het vertrouwen dat kandidaten ervaren als ze weten dat ze hun antwoorden mogen herzien en slordigheidsfoutjes kunnen herstellen (zie ook Bergstrom & Lunz, 1992).

In de *review*-conditie maakte 61% van de kandidaten gebruik van de mogelijkheid om antwoorden te herzien en zo nodig te veranderen. Daarbij werd 2% van de antwoorden daadwerkelijk veranderd. Item review leidde tot een kleine verbetering van de prestaties, alhoewel het verschil niet significant was. De meeste kandidaten die antwoorden veranderden, deden dat bij tussen de één en vier items. De conclusie was dat review van minimale invloed op de prestaties is.

In de *defer*-conditie maakte bijna de helft (45%) van de kandidaten gebruik van deze mogelijkheid. Daarbij werd bij 2.5% van de items de beantwoording uitgesteld tot aan het einde van het examen. Kandidaten bleken vooral moeilijke items uit te stellen. Geen van de toetsingen van de verschillen met de prestaties in de drie andere condities gaf significantie te zien.

Binnen elk van de condities waren er kandidaten die geen gebruik maakten van de mogelijkheid tot overslaan, uitstellen of herzien. Tussen de groep die dat wel en niet deed, waren geen verschillen in gemiddelde prestatie aantoonbaar en dat gold binnen elk van de drie condities.

Tot sluit wijzen we erop dat de resultaten van deze studie verkregen zijn met een CAT. In een CAT is item review om verschillende redenen veel problematischer dan in een lineaire CBT. De resultaten van Lunz et al. (1992) zijn dan ook beperkt generaliseerbaar naar lineaire CBT.

**Olea, Revuelta, Ximénez en Abad (2000)**

Olea, Revuelta, Ximénez en Abad (2000) vergeleken twee digitale versies (met en zonder item review) van een lineaire en een adaptieve Engelse woordenschattoets bij een steekproef van Spaanstalige eerste jaarstudenten psychologie. Van alle studenten veranderde 82% één of meer antwoorden waarbij 14% van de antwoorden gewijzigd werd. Het toestaan van item review resulteerde in een hoger percentage goede antwoorden en een hogere vaardigheid van de leerlingen, maar ook in een aanzienlijke toename van de afnametijd. De correlatie tussen de scores zonder en met item review bedroeg .94 voor de lineaire CBT en .95 voor de adaptieve CBT (CAT).

**Mason, Patry en Bernstein (2001)**

Mason, Patry en Bernstein (2001) vergeleken de scores op PPT en CBT met een vergelijkbare flexibiliteit. Dit wil zeggen: met vergelijkbare mogelijkheden om terug te keren naar eerder gemaakte vragen, vragen over te slaan, antwoorden opnieuw te bekijken en zo nodig te veranderen en dergelijke. Aan het onderzoek namen 27 psychologiestudenten deel. Voor de prestaties bleek het niet uit te maken of de studenten de PPT of de CBT maakten. De onderzoekers schrijven dit toe aan het gegeven dat hun CBT de flexibiliteit van een PPT nauwkeurig nabootste (en de motivatie van de studenten in beide condities even hoog was). De kandidaten konden op effectieve wijze door de CBT navigeren en behaalden daardoor met PPT equivalente scores.

**Russell, Goldberg en O'Connor (2003)**

Op basis van een review concluderen Russell et al. (2003) dat "*administration factors, such as transfer of problems from the screen to scratchwork space, lack of scratchwork space, and inability to review and/or skip individual test items, were found to affect [computer-based] test performance significantly*" (p. 282). Verder concluderen zij dat "*research on some mathematics tests indicates that validity is threatened when students experience difficulty accessing scratch paper in which they perform calculations*" (p. 288). Voor een bespreking van de uitkomsten van de door hen aangehaalde studies wordt de geïnteresseerde lezer verwezen naar paragraaf 3.4 en 3.5 van dit rapport.

**Revuelta, Ximénez en Olea (2003)**

Revuelta, Ximénez en Olea (2003) namen drie versies van een digitale test Engels voor Spaanstalige eerste-jaarstudenten psychologie. De drie versies waren lineaire CBT, adaptieve CAT en een CAT-versie (ECAT) met toewijzing van iets gemakkelijkere items dan bij de 'gewone' CAT. De vier condities waren: geen item review, review alleen aan het einde van de toets, review per blok van vijf items en review voor elk item.

Item review resulteerde bij alle drie versies in hogere prestaties, maar ook in een aanzienlijke verlenging van de toetstijd (gemiddeld over alle versies en condities met 21%). Van alle leerlingen maakte 90% gebruik van de mogelijkheden tot item review, waarvan 65% van item review profiteerden.

De onderzoekers rapporteren ook de uitkomsten afzonderlijk voor de lineaire CBT-versie (die het meest vergelijkbaar is met de rekentoets). Van de leerlingen maakte 87% gebruik van item review. Van alle antwoorden werd 16% gewijzigd. Van de gewijzigde antwoorden werd 59% gewijzigd van fout naar fout, 32% van fout naar goed en 9% van goed naar fout. Van alle leerlingen in de CBT-conditie profiteerde 64% van item review.

**Paek (2005)**

Paek (2005) reviewde een aantal K12-studies met meerkeuzetoetsen die vóór en na 1993 waren uitgevoerd. Zij concludeert dat de scores op recentere PPT en CBT vergelijkbaar zijn over leerjaren en vakgebieden. Zij geeft hiervoor twee mogelijke verklaringen. Een eerste verklaring is dat kandidaten tegenwoordig vaardiger zijn in het gebruik van de computer dan voorheen. Als tweede verklaring noemt zij het gegeven dat moderne toets-systemen vergelijkbare mogelijkheden bieden om vooruit en achteruit te bladeren, items over te slaan en itemantwoorden te bekijken en te veranderen als traditionele papieren toetsen. De nieuwe navigatietools maken het mogelijk dezelfde *test-taking-strategies* toe te passen als bij papieren toetsen, met als resultaat meer equivalente scores. Een uitzondering vormen meerkeuze-items met lange leesteksten waar kandidaten moeten scrollen: die

zijn digitaal nog steeds moeilijker dan op papier. Bij rekenen-wiskunde zijn lange teksten echter een zeldzaamheid, al komt het wel voor dat informatie verdeeld is over meer beeldschermpagina's en kandidaten dus moeten scrollen. Zie paragraaf 3.1 voor meer informatie over deze review.

**Leeson (2006)**

Leeson (2006) geeft een samenvatting van zeventig jaar onderzoek naar het effect van het al dan niet kunnen herzien van antwoorden op toetsvragen. Onder *item review* verstaat zij de mogelijkheid om items opnieuw te bekijken, over te slaan en gegeven antwoorden te veranderen. De resultaten laten volgens haar consistent zien dat de meerderheid van de kandidaten slechts een klein aantal antwoorden verandert, waardoor de toetsresultaten meestal verbeteren.
Leeson merkt op dat item review in de meeste lineaire CBT nog niet mogelijk is, maar dat er tegenwoordig goede algoritmen bestaan die structurele review mogelijk maken.
Leeson (2006) noemt het volgende onderzoek naar het effect van het toestaan van item review in lineaire CBT op de toetsprestaties:

- Van een meerkeuze-examen vergeleken Eaves en Smith (1986) een PPT-versie met mogelijkheden tot item review en een CBT-versie zonder deze mogelijkheden en vonden geen verschil in gemiddelde prestaties.
- Spray, Ackerman, Reckase en Carlson (1989) vergeleken een PPT met een CBT waarbij kandidaten vrij door de toets konden navigeren en waarbij zij eerdere antwoorden konden bekijken en veranderen en vonden geen verschil tussen de gemiddelden en de cumulatieve scoreverdelingen.
- In een soortgelijke studie vergeleken Luecht, Hadadi, Swanson en Case (1998) een gewone PPT met twee CBT's: één met ingebouwde flexibiliteit en één zonder. Het ontbreken van flexibiliteit had geen effect op de prestaties. Wel bleken kandidaten het gebrek aan flexibiliteit maar matig te kunnen waarderen.
- Vispoel (2000) bestudeerde de antwoorden van kandidaten op een woordenschattoets voorafgaand en nadat zij de mogelijkheid tot item review hadden gekregen en vond een positief effect op de prestaties (zie eerder in deze tekst voor een korte bespreking van dit onderzoek).

## 3.4 Gebruik van kladpapier

**Inleiding**

Examens rekenen-wiskunde bevatten vaak veel vragen waarbij kandidaten iets moeten uitrekenen. Het antwoord vereist doorgaans meerdere denkstappen of oplossingsprocessen waarvan de kandidaat er één of meer goed uitgevoerd kan hebben. Bij sommige typen opgaven kan of moet de kandidaat het daartoe benodigde rekenwerk 'uit het hoofd' uitvoeren, bijvoorbeeld als de opgave erg eenvoudig is, als er iets geschat moet worden of wanneer het hoofdrekenenopgaven betreft. Bij deze opgaven is een effect van de afnamemodus minder waarschijnlijk.
Bij de meer ingewikkelde opgaven zijn tussenberekeningen, uitwerkingen en dergelijke noodzakelijk die kandidaten (nog) niet op de computer kunnen uitvoeren en documenteren. De kandidaat staat dan voor de keuze: doe ik het uit het hoofd of gebruik ik kladpapier (Engels: *scratch paper*, *scrap paper*, *rough working*)? Het kladpapier bevat als het ware de gedocumenteerde evidentie van de gehanteerde strategieën en denkprocessen (Johnson & Green, 2006).

Er zijn verschillen tussen het gebruik van kladpapier in een CBT en een PPT die implicaties kunnen hebben voor de vergelijkbaarheid van beide afnamemodi. Achter de computer moeten leerlingen als zij kladpapier gebruiken telkens schakelen tussen beeldscherm en papier, terwijl zij bij de papieren versie hun uitwerkingen in de kantlijn van het toetsboekje kunnen noteren; bij een PPT werkt de leerling in een twee-dimensioneel vlak, terwijl een CBT het veelvuldig switchen tussen beeldscherm en kladpapier in drie dimensies vereist (Kingston, 2009). Daardoor zouden kandidaten bij CBT minder geneigd zijn om kladpapier te gebruiken dan bij PPT, met als gevolg een grotere kans op fouten en lagere prestaties (Commissie Bosker, 2014).

Er zijn aanwijzingen dat het niet gebruiken van kladpapier een negatief effect op de prestaties heeft. Zo laten periodieke peilingen van het onderwijsniveau in Nederland zien dat de prestaties op het onderdeel 'bewerkingen' achteruit zijn gaan. Als belangrijkste oorzaak wordt genoemd dat veel leerlingen tegenwoordig ten onrechte geen kladpapier gebruiken, maar proberen de opgaven 'uit het hoofd' te maken (Janssen, Van der Schoot & Hemker, 2004; Van Putten, 2008; Scheltens, Hickendorff, Eggen & Hiddink, 2014; Hickendorff, Heiser, Van Putten & Verhelst, 2009).

Recente vragenlijstgegevens (Cito, 2015a) laten zien dat ruim twee derde (69%) van de kandidaten die deelnamen aan de Rekentoets VO naar eigen zeggen bij de meeste opgaven kladpapier gebruikte, ruim een kwart (28%) deed dat af en toe en 3% bij geen enkele opgave.

In het vervolg van deze paragraaf bespreken we eerst enkele reviews waarin het gebruik van kladpapier ter sprake komt en vervolgens enkele individuele studies naar het verschillend gebruik van kladpapier in PPT en CBT en het effect ervan op de prestaties.

**Resultaten: reviews en meta-analyses**

In een review vonden Lee en Hopkins (1985) hogere gemiddelde scores op PPT in vergelijking met CBT (zie ook paragraaf 3.1). Als belangrijke oorzaak noemen zij de beperkte ruimte voor het werken met kladpapier en het frequent moeten wisselen tussen scherm en kladpapier.
Eveneens op basis van een review concluderen Russell, Goldberg en O'Connor (2003) dat "*administration factors, such as transfer of problems from the screen to scratchwork space, lack of scratchwork space (….) were found to affect [computer-based] test performance significantly*" (p. 282). Ook concluderen zij dat "*research on some mathematics tests indicates that validity is threatened when students experience difficulty accessing scratch paper in which they perform calculations*" (p. 288).
In een meta-analyse analyseerde Kingston (2009) onder meer de resultaten van 31 vergelijkingsstudies op het gebied van wiskunde. PPT wiskunde werden gemiddeld iets beter gemaakt dan CBT wiskunde. De gemiddelde effectgrootte was met -.06 echter zeer klein (met een 95%-betrouwbaarheidsinterval van -.10 tot -.02). Kingston noemt het waarschijnlijk dat dit kleine verschil te maken heeft met het verschillend gebruik van kladpapier. Kingston spreekt de hoop uit dat toetsontwikkelaars manieren bedenken om dit probleem te omzeilen, net zoals zij dat bij het lezen van grote hoeveelheden tekst gedaan hebben (bijvoorbeeld door leerlingen in de gelegenheid te stellen eerder gelezen tekst digitaal te markeren en te annoteren).

**Resultaten: individuele studies**

**Braswell en Bridgeman (1992)**

De studie van Braswell en Bridgeman (1992) betreft weliswaar niet het verschillend gebruik van kladpapier in PPT en CBT en het effect ervan op de prestaties, maar we vermelden de resultaten hier toch omdat het inzicht verschaft in het effect van de fysieke afstand tussen de toetsvraag en de ruimte voor uitwerkingen. De onderzoekers vergeleken twee papieren 25-itemversies van de SAT-Mathematical (SAT-M): één in het standaardformat en één in datzelfde format met een extra kolom voor kladwerk. Zij vonden geen effect op de prestaties. In beide formats maakte de meerderheid van de kandidaten bij minimaal driekwart van de items gebruik van kladwerk. Ogenschijnlijk eenvoudige berekeningen, zoals 1000 x 10, werden daarbij vaak uitgeschreven. Echter, zelfs als voor het kladwerk een aparte kolom in het toetsboekje was gereserveerd, gaven de meeste kandidaten er de voorkeur aan het kladwerk bij het item zelf te noteren in plaats van op een plek verder weg van het item.

**Hickendorff, Van Putten, Verhelst en Heiser (2010)**

De studie van Hickendorff, Van Putten, Verhelst en Heiser (2010) is eveneens niet volledig on-topic, maar we maken er toch melding van omdat het inzicht geeft in het effect van strategiekeuze (mentaal versus schriftelijk) in relatie tot achtergrondkenmerken van de leerlingen.
Hickendorff, Van Putten, Verhelst en Heiser (2010) toonden aan dat het vooral zwakke rekenaars en jongens zijn die bij complexe deelsommen kiezen voor zuiver mentale strategieën in plaats van kladpapier te gebruiken (en zichzelf daarmee benadelen). Als men leerlingen die 'uit het hoofd' een som uitrekenen, dwingt om kladpapier te gebruiken bij een parallelle opgave, komt dat hun prestaties ten goede. Vandaar dat de onderzoekers de aanbeveling doen het gebruik van kladpapier bij complexe deelproblemen te stimuleren.

**Van den Heuvel-Panhuizen en Bodin-Baarends (2004)**

Van den Heuvel-Panhuizen en Bodin-Baarends (2004) namen een papieren rekentoets af bij hoog-presterende leerlingen uit het Nederlandse basisonderwijs. De toets bevatte vijftien puzzel-achtige opgaven zoals *number riddles.* De opgaven bevatten grote hoeveelheden gegevens waarbij de leerling de oplossing vaak kon vinden door systematisch opties uit te proberen. Een van de uitkomsten was dat deze leerlingen zelden gebruik maakten van de ruimte voor kladwerk, waardoor zij, aldus de onderzoekers, lager dan verwacht presteerden.

**Gaskill en Marshall (2006)**

Gaskill en Marshall (2006) deden onderzoek naar de vergelijkbaarheid van PPT en CBT bij meerkeuzetoetsen *numeracy* in grade 4 en 7. Bij beide versies had de leerling papier nodig om het antwoord te kunnen geven. Bij de PPT-versie konden de leerlingen hun uitwerkingen en berekeningen op dezelfde pagina en vaak op dezelfde regel noteren als het item. Wel moesten zij het antwoord overbrengen naar een antwoordblad dat zich echter in de onmiddellijke nabijheid van het toetsboekje bevond. Bij de CBT-versie moesten de leerlingen de benodigde informatie eerst van het scherm overbrengen naar het kladpapier, vervolgens de oplossing op het kladpapier uitwerken en tot slot het resultaat weer 'terugbrengen' naar het scherm. Bij de CBT-versie is de kans op het maken van fouten dus groter dan bij de PPT-versie. De veronderstelling was dan ook dat de CBT-versie tot lagere scores zou leiden dan de PPT-versie en dat het verschil groter zou zijn voor de laag dan voor de hoog vaardige leerlingen. Leerlingen behaalden inderdaad hogere scores op de PPT-versie van de

meerkeuzetoets over rekenvaardigheid dan op de CBT-versie maar het verschil was niet groter voor de minder vaardige leerlingen. Gaskill en Marshall (2006) bevelen scholen aan de digitale afnamesituatie zo in te richten dat de leerlingen voldoende ruimte hebben om de berekeningen op papier uit te voeren en ervoor te zorgen dat zij de informatie van het scherm gemakkelijk op het kladpapier kunnen overnemen (en vice versa). Verder wijzen de onderzoekers op het belang van een goede voorbereiding van de leerlingen op de digitale toetsing. Zo zouden leerlingen getraind moeten worden in het overnemen en controleren van informatie van het scherm naar papier (en vice versa).

**Johnson en Green (2006)**

Johnson en Green (2006) vergeleken PPT- en CBT-versies van dezelfde rekentoets bij elf-jarigen in het Engelse primair onderwijs. De leerlingen werd gevraagd hun tussenberekeningen en dergelijke zo uitgebreid mogelijk te documenteren in de daarvoor bestemde ruimte op het toetsboekje (PPT) of op het afzonderlijk ter beschikking gestelde kladpapier (CBT). Daarnaast namen de onderzoekers de leerlingen interviews af. Daarbij werd gevraagd naar verschillen tussen beide versies van een opgave in de manier waarop zij het probleem hadden uitgewerkt. Bij de CBT-versie werd daarbij ook gebruik gemaakt van de *Replay Option* waarmee de leerlingen hun antwoorden en revisies in chronologische volgorde konden terugzien. Verder werden vragen gesteld over hun voorkeur voor vraagtypen in de ene en andere afnamemodus. Met behulp van inhoudsanalyse codeerden de onderzoekers de gemaakte fouten aan de hand van de uitwerkingen op papier. Op basis van argumenten zoals genoemd in de inleiding op deze paragraaf, veronderstelden zij dat leerlingen bij de CBT-versie minder geneigd zouden zijn tot verstrekken van evidentie over de gehanteerde strategieën en denkprocessen dan bij de PPT-versie.

**Kwantitatieve resultaten**

Johnson en Green (2006) vonden dat de PPT-versie iets makkelijker was dan de CBT-versie, maar het verschil tussen de gemiddelde totaalscores was niet significant. Een analyse op itemniveau bracht aan het licht dat van de zestien vragen er elf significant gemakkelijker waren in de PPT-versie; slechts één vraag was significant gemakkelijker in de CBT-versie.

De meest gemaakte fouten waren *computation and mental calculation errors* die leerlingen maakten als uitgeschreven uitwerkingen en dergelijke ontbraken en zij kennelijk geprobeerd hadden de vraag uit het hoofd te beantwoorden. In de CBT-versie werden deze mentale rekenfouten vaker gemaakt dan in de PPT-versie.

De verschillen tussen de beide afnamemodi in het aantal gemaakte rekenfouten bleken afhankelijk van de aard van de vraag. Bij alle vragen die een beroep deden op aftrekken met *substraction using decomposition*, bijvoorbeeld 554-538 of 546-39, maakten de leerlingen meer mentale rekenfouten in de digitale dan de papieren versie. Verder werden in de CBT-versies van lange vermenigvuldigingsvragen meer opsplitsfouten (Engels: *partitioning errors*) gemaakt dan in de PPT-versies (dit wil zeggen: het splitsten van grote gehele getallen in kleine gehele getallen voorafgaand aan een rekenkundige bewerking (Engels: operation). Johnson en Green (2006) interpreteerden dit als volgt: "*This meant that students made more errors when separating out the Tens and Units components of large numbers, and tended to have more problems multiplying the appropriate parts when working on the computer*." (p. 16).

Er werden door de bank genomen relatief weinig transcriptiefouten gemaakt, dit wil zeggen fouten bij het overnemen van gegevens binnen dezelfde pagina of van de ene naar de andere pagina (PPT) of van het scherm naar kladpapier en vice versa (CBTP). Echter, bij de CBT-versie werden zoals verwacht meer transcriptiefouten gemaakt dan bij de PPT-versie. Transcriptieproblemen van scherm naar papier en omgekeerd deden zich voor bij ongeveer tien procent van de leerlingen. Johnson en Green (2006) veronderstellen dat het grotere aantal transcriptiefouten samenhangt met de grotere fysieke afstand die de informatie moet overbruggen gedurende de verwerking van het probleem. Bij een papieren versie zijn de locaties van de vraag, uitwerking en antwoorden in elkaars onmiddellijke nabijheid. Bij een CBT is de fysieke afstand en daarmee de belasting van het werkgeheugen daarentegen veel groter. Eerst moet de leerling de vraag op het scherm lezen, vervolgens moet hij of zij deze informatie vasthouden in het werkgeheugen als de aandacht verschuift naar het kladpapier en vervolgens weer naar het scherm, om ten slotte het antwoord in het daarvoor bestemde invoervak op het scherm in te typen. Dat leerlingen in de CBT-versies meer transcriptiefouten maken, betekent volgens Johnson en Green (2006) dat "*that their lack of success should not have been attributed to them having conceptual problems relating to the particular question within which the error was found*" (p. 24). Deze bevinding roept volgens de onderzoekers vragen op over de validiteit van scores zoals verkregen met digitale toets- en examensystemen die diagnostische profielen rapporteren op basis van foutenanalyses. Hun advies is na te denken over de manier waarop het aantal transcriptiefouten verder kan worden teruggebracht, bijvoorbeeld door leerlingen in staat te stellen annotaties op het scherm te maken.

Op papier werden meer vragen niet beantwoord dan op het scherm, waarbij twee keer zoveel meisjes dan jongens een of meer vragen onbeantwoord lieten. Waarschijnlijk vanwege het gebrekkige onderscheiden vermogen - het aantal leerlingen was klein - was dit verschil echter niet significant. Ter verklaring verwijzen de auteurs naar Gallagher, Bridgmen en Cahalan (2000) die veronderstellen dat CBT voor sommige leerlingen een minder bedreigende omgeving vertegenwoordigt dan PPT. Johnson en Green (2006) veronderstellen dat de jongens in hun onderzoek meer geneigd zijn *to take a chance* dan meisjes, ook al zijn zij er niet zeker van dat hun antwoord goed is. Een mogelijke verklaring is dat leerlingen het beantwoorden van digitale vragen associëren met andere activiteiten die zij op de computer uitvoeren, zoals spelletjes, waarbij wordt uitgegaan van de filosofie van *have a go and start again* (Gallagher et al., 2000) of *snatch and grab* (Sutherland-Smith (2002).
Daarnaast zouden de verschillen tussen de beide afnamemodi in het onbeantwoord laten van vragen ermee te maken kunnen hebben dat het indienen van antwoorden via het beeldscherm als een minder persoonlijke aangelegenheid wordt beschouwd dan het toevertrouwen van antwoorden aan het papier. Johnson en Green (2006) lichten dit als volgt toe: "*When students answer on paper their attempts and errors are made explicit and public, whereas the computer creates a more private workspace where students may be more willing to risk being wrong. When answers are submitted on-line there is no immediately visible trace of evidence relating to past questions which the student may have struggled with, and that they need to confront each time that they look at any subsequent question, although this information might be stored elsewhere for teacher analysis at a later time. This contrasts with the paper versions of each test, which expose a student's prior attempts at answers in the public arena occupied by themselves, and potentially their peers and teachers. Having the opportunity to submit answers in a less public environment may lead students to worry less about*

*the type of answers that they give, perhaps encouraging them to take risks about which strategies to employ*" (p. 28).

De onderzoekers vonden ook modus-gerelateerde verschillen in de mate waarin leerlingen hun werkwijze bij de vraag documenteerden. Bij negen van de zestien vragen gaven de leerlingen bij de CBT-versie meer informatie over hun werkwijze dan bij de PPT-versie, bij vier vragen werd de werkwijze bij de PPT-versie beter gedocumenteerd en bij drie vragen was er geen verschil.

**Kwalitatieve resultaten**

Drie vragen werden op papier significant beter gemaakt dan via het beeldscherm. Een kwalitatieve analyse bracht aan het licht dat het drie van de vier vragen betrof vragen waarbij meer leerlingen hun uitwerkingen aan het papier toevertrouwden bij de PPTT- dan de CBT-versie. De onderzoekers veronderstellen dat de extra inspanning die nodig is om het denkproces in de digitale afnamemodus op papier te documenteren ertoe geleid heeft dat deze leerlingen de berekeningen uit het hoofd uitvoeren (in plaats van kladpapier te gebruiken). De foutenanalyse bood verdere ondersteuning voor deze veronderstelling. Bij deze drie vragen maakten de leerlingen namelijk meer *combined computational and mental errors* in de CBT- dan in de PPT-versie. Bij deze drie vragen zou de aversie tegen het gebruik van kladpapier er dus toe leiden dat leerlingen meer op zuiver mentale strategieën vertrouwen waardoor zij meer fouten maken en lagere prestaties behalen.
Een kwalitatieve analyse liet verder zien dat 37% van de leerlingen bij de papieren en digitale versie verschillende oplossingsstrategieën toepasten. Zo waren leerlingen bij de papieren versie van vragen zoals '554-538' en '546-30' meer geneigd om *partitioning strategies* te gebruiken dan bij de digitale versie. De algehele conclusie was dat leerlingen bij de PPT-versie wat vaker informele en flexibele oplossingsstrategieën hanteerden dan bij de CBT-versie waar vaker formele en standaardstrategieën werden toegepast. De groep die flexibele oplossingsstrategieën hanteerde, bevatte overigens significant meer meisjes dan jongens.

**Threfall, Pool, Homer en Swinnerton (2007)**

Threfall, Pool, Homer en Swinnerton (2007) deden vergelijkingsonderzoek bij leerlingen van 11 en 14 jaar in Groot-Brittannië. Bij elfjarigen waren de totaalscores voor de 24 items bij de CBT-versie 3% hoger dan bij de PPT-versie en bij viertienjarigen scoorde de CBT 5% lager dan de PPT. Hoewel Threfall et al. (2007) geen gegevens over de statistische significantie verstrekken, concludeerden zij toch dat "*A difference of 5% or less in performance cannot be said to be indicative of an underlying effect*" (p. 340).
Zeven items waarbij het verschil in p-waarde tussen CBT en PPT tussen de 12% en 34% bedroeg, werden nader bestudeerd. Bij vijf ervan scoorde de CBT hoger dan de PPT. Bij vier van deze vijf items moest de leerling elementen in de goede volgorde zetten. Een mogelijke verklaring veronderstelt dat het (op interactieve wijze) exploreren van de verschillende arrangementen op de computer minder belastend is voor het geheugen dan op papier.
Op twee van de zeven items scoorde PPT hoger dan CBT. Bij beide items hadden de leerlingen kladpapier nodig om het juiste antwoord te kunnen geven. Hoewel de leerlingen een *working booklet* tot hun beschikking hadden, werd dat door maar weinig leerlingen gebruikt. Anders dan in het onderzoek van Johson and Green (2006) het geval was, gebruikten naar verhouding meer leerlingen kladpapier bij de PPT- dan bij de CBT-versie.

**Keng, McClarty en Davis (2008)**

Keng, McClarty en Davis (2008) vergeleken een PPT- en een CBT-versie van onder meer een wiskundetoets voor grade 8 en 11. Hoewel de toets ook open vragen bevatte, gebruikten de onderzoekers alleen de meerkeuzevragen. Beide versies boden mogelijkheden tot item review. In de papieren versie konden de leerlingen hun aantekeningen en tekeningen direct in het toetsboekje maken. In de online versie kon de leerling gebruik maken van online *drawing tools* zoals een *highlighter* en een *answer eliminator*.

Bij vier van de vijftig wiskunde-items voor grade 8 werd een effect van de afnamemodus gevonden waarvan drie in het voordeel van de PPT-versie. Bij twee van deze drie items moest de leerling grafische en geometrische manipulaties uitvoeren. Ter verklaring voeren Keng et al. (2008) aan dat digitaal tekenen moeilijker is dan tekenen op papier. Het overzetten van de grafiekjes van beeldscherm naar papier is een extra stap die bij de papieren versie niet nodig is. Het overzetten van grafiekjes vereist een zekere nauwkeurigheid waardoor de CBT-versie van het item een (onbedoeld) extra beroep zou kunnen doen op de tekenvaardigheid van de leerling. Het derde item waarbij de PPT-versie tot hogere scores leidde, kende verschillende omvangrijke *graphics* of geometrische vormen waarbij de leerling moest scrollen. Keng et al. (2008) veronderstellen de CBT-versie van dat item moeilijker was omdat de leerlingen om het hele item te kunnen bekijken voortdurend moesten scrollen, terwijl ze het volledige item in hun toetsboekje op een en dezelfde pagina konden bekijken. De onderzoekers vonden geen verklaring waarom het vierde item het in de CBT-conditie beter deed dan in de PPT-conditie.

De resultaten in grade 11 gaven een vergelijkbaar beeld te zien. Ook daar werden items die grafische en geometrische manipulaties en scrollen vereisten in het PPT-format beter gemaakt dan in het CBT-format. De onderzoekers vermelden dat hun resultaten bij wiskunde overeenkomen met die van Sandene et al. (2005) en Greenwood et al. (2000). Laatstgenoemden vonden dat items die een beroep doen op *spatial and gross motor skills* digitaal moeilijker zijn dan op papier.

## 3.5 Invoer van antwoorden via beeldscherm versus papier

**Inleiding**

Onderzoek laat zien dat verschillen in ICT-ervaring mede verantwoordelijk kunnen zijn voor prestatieverschillen tussen PPT en CBT (o.a. Gaskill & Marshall, 2006). Als een mogelijke verklaring voor de tegenvallende resultaten op de rekentoetsen noemt de Commissie Bosker (2014) het gegeven dat kandidaten over het algemeen onbekend zijn met het digitaal afnemen van toetsen en examens. Een specifiek probleem bij CBT rekenen-wiskunde is dat de computer specifieke eisen stelt aan de invoer van de antwoorden. Kandidaten zijn daar vaak nog onvoldoende van op de hoogte en in getraind. Velen hebben meer ervaring met het noteren van antwoorden op papier dan met het invoeren ervan via het beeldscherm (McGuire & Youngson, 2002). In een CBT voeren de kandidaten hun antwoord uiteindelijk in op het scherm, maar voor de berekeningen en uitwerkingen op kladpapier moeten zij toch nog potlood en papier gebruiken. Als zij daarmee klaar zijn, moeten zij hun ‘papieren’ antwoord vertalen in wiskundige symbolen die de computer kan herkennen en dat antwoord vervolgens transponeren naar het beeldscherm. Onder meer door een gebrek aan ervaring met de invoermodule zouden kandidaten bij een CBT meer invoerfouten maken dan bij een PPT, met als gevolg lagere prestaties voor CBT in vergelijking met PPT. De afnamemodus introduceert hier constructirrelevante variantie die afbreuk kan doen aan de validiteit.

Een hiermee samenhangend probleem van CBT-examens rekenen-wiskunde heeft te maken met de bezorgdheid van kandidaten dat de scoringsprogrammatuur onvoldoende recht doet aan hun kennis en vaardigheden (o.a. McGuire & Johnson, 2002). In een examen rekenen-wiskunde moeten kandidaten doorgaans iets uitrekenen en hun antwoord in de vorm van een getal of een formule aan het papier toevertrouwen of in de computer invoeren. Bij CBT-examinering van rekenen-wiskunde maken kandidaten zich zorgen over de volgende problemen die niet of in mindere mate optreden bij menselijke beoordeling van antwoorden op papier (McGuire & Johnson, 2002):

- de computer interpreteert een ingevoerd getal of een ingevoerde formule op een andere manier dan de kandidaat bedoeld heeft;
- de computer rekent een goed antwoord fout omdat het in het verkeerde format gegeven is;
- de computer herkent een juiste numerieke benadering niet als een goed antwoord;
- er zijn meer goede antwoorden mogelijk, maar de computer herkent er slechts één als juist;
- de computer rekent gedeeltelijk goede antwoorden helemaal fout.

**Resultaten: reviews en meta-analyses**

**Mazzeo and Harvey (1988)**

In hun review verwijzen Mazzeo and Harvey (1988) naar onderzoek waarbij de door de computer gelezen antwoorden werden vergeleken met de antwoorden op papier. De computer bleek lang niet alle ingevoerde antwoorden correct te lezen. Ook Bunderson et al. (1989) wijzen er in hun review op dat de relatief hoge scores op PPT mede veroorzaakt worden doordat de computer juiste antwoorden van kandidaten ten onrechte als fout interpreteert. Hierna bespreken we de resultaten van enkele studies waarin geprobeerd is de invoerproblematiek van CBT te minimaliseren.

**Resultaten: individuele studies**

**Beevers, McGuire, Stirling en Wild (1995)**

In een van de eerste experimenten met geautomatiseerde partial credit scoring in een CBT wiskunde via de Steps-methode (zie paragraaf 3.2) probeerden Beevers et al. (1995) ook een oplossing te vinden voor het probleem van wat zij de mismatch tussen antwoorden op het beeldscherm versus op papier noemen. Aanleiding tot deze deelstudie was dat leerlingen vaak bezorgd zijn over het gegeven dat zij het invoeren van hun antwoorden syntactische fouten maken en dan niet in de gaten hebben dat de computer hun antwoord verkeerd interpreteert. Het probleem van de mismatch tussen antwoorden op het beeldscherm versus op paper werd aangepakt door een *Input Tool* in de CBT in te bouwen. Deze optionele tool liet de ingevoerde input - bijvoorbeeld een antwoord met een breuk of machtsverheffen - opnieuw zien, maar dan in een wiskundige notatie die vergelijkbaar was met de manier waarop de student de input op papier zou hebben genoteerd. Zo vereiste het scoringsalgoritme het plaatsten van haakjes rond het argument van functies (Engels: *the argument of function*s). Het juiste antwoord ‘de sinus van 2x’, dat een menselijke beoordelaar (gedeeltelijk) goed zou rekenen, moest bijvoorbeeld worden ingetypt als ‘sin(2x)’ en niet als ‘sin 2x’. Om dit probleem op te lossen, gaf de *Input Tool* dan de foutmelding dat de uitdrukking verkeerd geformuleerd was. Vrijwel alle studenten bleken de *Input Tool* te gebruiken, waardoor hun bezorgdheid aanzienlijk afnam.

**Beevers, Youngson, McGuire, Wild en Fiddes (1999)**

In de loop der jaren is de *Input Tool* op basis van onderzoek bij kandidaten aanzienlijk verbeterd. Misinterpretaties tussen kandidaat en computer zijn verder geminimaliseerd. In een vervolgpublicatie beschrijven Beevers et al. (1999) een *syntax checker* die kandidaten ondersteunt bij het vormen van betekenisvolle wiskundige uitdrukkingen (Engels: *meaningful expressions*). De tool bevat ook een dynamisch display die de kandidaat onmiddellijk feedback geeft over hoe de computer het ingevoerde antwoord interpreteert. Als het antwoord dicht tegen het goede antwoord aanzit maar niet de vereiste vorm heeft, kan de leerling toch deelscores verdienen. Hoe dat in zijn werk gaat, wordt duidelijk uit het volgende voorbeeld (zie voor meer voorbeelden: Beevers et al. , 1999). Als het toegestane antwoord op een vraag ½ is maar de leerling heeft 0.5 ingetypt, geeft de computer - als de leerling de *Input Tool* heeft aangezet - de boodschap '*Je antwoord is goed, maar heeft de verkeerde vorm; geef je antwoord als een breuk - 50% aftrek'*. Leerlingen blijken de toegenomen flexibiliteit bijzonder op prijs te stellen, terwijl hun toetsangst daardoor afneemt.

**Ashton, Beevers, Korabinski, Youngson en Martin (2006)**

Een van de doelen van het onderzoek van Ashton, Beevers, Korabinski, Youngson en Martin (2006) was het vinden van een manier om invoerfouten in een CBT te beperken door de kandidaat onmiddellijke feedback te geven op het ingevoerde antwoord. Aanleiding was de observatie dat veel foute antwoorden in een traditionele CBT het gevolg zijn van kleine reken- of slordigheidsfoutjes die in een PPT nauwelijks tot puntenaftrek geleid zouden hebben. Een voorbeeld is het op zich juiste antwoord *twenty one and a third* waarbij de kandidaten niet altijd wisten hoe zij de breuk (Engels: *fractional part*) moesten invoeren opdat de computer deze zou herkennen. Door de kandidaat hierop te attenderen, zou deze de kans krijgen om de fout te traceren en het foute antwoord door het goede te vervangen.

De onderzoekers ontwikkelden een zogenoemde T-versie van het examen waarbij de feedback werd gegeven in de vorm van vinkjes en kruisjes (Engels: *ticks and crosses*) voor respectievelijk een goed en een fout antwoord. De gemiddelde scores op deze T-versie verschilde niet van die op de beide andere versies: *Compulsory Steps (CS)* en *Optional Steps* (OS). Zie de bespreking van de resultaten met betrekking tot het effect van het al dan niet toekennen van *partial credit* in paragraaf 3.2.

De onderzoekers bestudeerden ook het aantal verbeterpogingen bij de T-versie waarbij de kandidaten feedback kregen in de vorm van vinkjes en kruisjes. De uitkomsten gaven aldus de onderzoekers aanleiding tot twijfel aan de validiteit van de meting met de T-versie. Het was de bedoeling dat de feedback kandidaten zou attenderen op kleine rekenfoutjes die zij dan vervolgens konden herstellen. In de T-versie bleken de kandidaten nog steeds rekenfoutjes te maken, ook al waren dat er minder dan bij de S-versie (CBT met steps). Wel roepen de resultaten volgens de onderzoekers de vraag op wat er in hoofden van de kandidaten omging als zij met een kruisje geconfronteerd werden en besloten door te gaan met de volgende vraag of het volgende onderdeel van de vraag. Anders dan de bedoeling was, waren er kandidaten die zeer veel pogingen ondernamen om een deel van de vraag te verbeteren. De onderzoekers achten het aannemelijk dat deze kandidaten gokgedrag vertoonden in de zin dat zij herhaaldelijk probeerden het antwoord te verbeteren zonder inzicht in de aard van de fout. Zij opperen dat het geven van feedback op de juistheid van het antwoord mogelijk meer geschikt is voor formatieve dan voor summatieve evaluatie.

Voor de onderzoekers waren de resultaten met de T-versie aanleiding om een aantal verbeteringen in het examen aan te brengen. Een verbetering was dat kandidaat, na het antwoord ingevoerd te hebben, twee weergaves van zijn of haar antwoord te zien krijgt (gepositioneerd onder het invoervak waarin zij hun antwoord noteerden). De eerste weergave laat het antwoord zien in dezelfde vorm als waarin zij het hebben ingetypt. De tweede weergave laat het antwoord zien zoals de kandidaat dat normaal gesproken op papier zou hebben gegeven. Deze zogenoemde *rendered form* is veel meer vertrouwd en gebruikelijk dan wat zij op het beeldscherm op één regel moeten intypen. De tweede weergave laat zien hoe de computer het antwoord geïnterpreteerd heeft. En als die interpretatie afwijkt van wat de kandidaat had willen invoeren, kan hij of zij het antwoord gemakkelijker verbeteren.
Als tweede verbetering stellen de onderzoekers voor het maximale aantal verbeterpogingen te beperken tot bijvoorbeeld maximaal twee. Zij wijzen erop dat dit tot nieuwe problemen kan leiden. Het komt bijvoorbeeld vaak voor dat kandidaten per ongeluk twee keer op een icoontje klikken wat dan als twee pogingen zou tellen. Hier is nader onderzoek geboden. De onderzoekers verwachten de geconstateerde invoerproblemen in de toekomst grotendeels te kunnen oplossen. Echter, net als er bij een PPT altijd misprints zullen voorkomen, zullen invoerfouten bij een CBT nooit helemaal vermeden kunnen worden.

**Passmore, Brookshaw en Butler (2011)**
Passmore, Brookshaw en Butler (2011) ontwikkelden een online systeem voor het toetsen van de vaardigheden van eerstejaars studenten in 'algebra and calculus'. Om de antwoorden - algebraïsche input - in te voeren, hoefden de studenten niet eerst een programmeertaal te leren of ingewikkelde invoerinstructies te bestuderen. De studenten gaven hun antwoorden volgens de gebruikelijke wiskundige notatiewijzen en conventies. Passmore, Brookshaw en Butler (2011) ontwikkelden een zogenoemd computer algebra system (CAS) waarmee de antwoorden werden vergeleken met bekende antwoorden, de input zo nodig ontleed werd om de kenmerken ervan nader te bepalen en de antwoorden automatisch gescoord werden. De invoer via het toetsenbord gebeurde in grote lijnen volgens de informele syntax van Sangwin en Ramsden (2007) die, aldus de onderzoekers, de gebruikelijke wiskunde notitiewijzen en conventies dicht benaderde. Het CAS gaf feedback aan de hand waarvan de student het antwoord kon controleren alvorens het definitief te versturen. Een voorbeeld van een syntactische hint is: "*A simple solution of the form x/y is expected. If you don't supply a simple solution it will be marked incorrect*" (p. 809). Het systeem voorzag ook in *partial credit scoring* (zie ook paragraaf 3.2).
Als voordelen noemen de auteurs de efficiëntie en de snelle feedback voor studenten en onderwijspersoneel. Studenten hadden geen klachten over het systeem, maar sommigen waren toch nog bang dat de computer hun antwoord vanwege een syntaxprobleem fout zou rekenen (zelfs als ander onderzoek uitwees dat zij een wiskundige fout gemaakt hadden). Passmore, Brookshaw en Butler (2011) melden dat het percentage syntactische fouten werd teruggebracht tot ongeveer 5% en dat dit minder is dan de 18% uit het onderzoek van Sangwin en Ramsden (2007).
De onderzoekers rapporteren ook enkele problemen, zoals het ontwikkelen en uittesten van goede vragen, het specificeren van de bijbehorende informele syntax en het gegeven dat "*no system can handle perfectly all the possible answers that students may give to any question*" (p. 902).

Het onderzoek van Passmore, Brookshaw en Butler (2011) is slechts één van de vele ontwerponderzoeken waarbij geprobeerd is programmatuur te ontwikkelen die overweg kan met

antwoorden op andere vragen dan meerkeuzevragen en dan vragen waarbij een exacte match met het gegeven numerieke antwoord vereist is. In het kader van onze kortdurende studie is het helaas niet mogelijk om de vele uitgevoerde ontwerponderzoeken te traceren, te bestuderen en samen te vatten.

# 4 Conclusies

In het voorgaande is verslag gedaan van de uitkomsten van een kortdurende literatuurstudie naar het effect van de afnamemodus (papier versus digitaal) op de hoogte van de prestaties op onder meer reken-wiskundetoetsen. Er blijken vele honderden studies te zijn uitgevoerd waarin de prestaties op PPT en CBT met elkaar worden vergeleken. De resultaten van dit onderzoek blijken sterk te variëren al naar gelang het vakgebied, de specifieke toets of examen, het vraagtype, het afnamesysteem en kenmerken van de populatie. De overgrote meerderheid van de studies naar rekenen-wiskunde zijn uitgevoerd met uitsluitend meerkeuzetoetsen (in *low-stakes* settings). Vergelijkingsstudies naar mogelijk differentiële effecten van *partial credit* scoring, het gebruik van kladpapier en de invoer van de antwoorden (beeldscherm versus papier) zijn nog nauwelijks uitgevoerd. De methodologische zuiverheid van deze studies laat bovendien vaak sterk te wensen over. Met dit in ons achterhoofd vatten we hieronder de belangrijkste conclusies van de literatuurstudie samen.

**Het effect van de afnamemodus op de prestaties**

De eerste onderzoeksvraag luidde: "Wat is er bekend over het effect van de afnamemodus op de hoogte van de prestaties van de kandidaten?". Omdat het aantal uitgevoerde studies zeer groot was, hebben we geprobeerd deze vraag te beantwoorden aan de hand van een beperkt aantal verhalende reviews en meta-analyses (Mazzeo & Harvey, 1988; Bunderson, Inouye & Olsen, 1989; Wise & Plake, 1989; Bergstrom, 1992; Dillon, 1992; Mead en Drasgow, 1993; Kim, 1999; Russell, Goldberg & O'Connor, 2003; Paek, 2005; Gaskill & Marshall; 2006; Wang, Jiao, Young, Brooks & Olson, 2007; Texas Education Agency, 2008; Kingston, 2009). De belangrijkste conclusies met betrekking tot de eerste onderzoeksvraag kunnen als volgt worden samengevat:

- de uitkomsten bij het vakgebied wiskunde geven weinig aanleiding om een sterk effect van de afnamemodus op de prestaties te veronderstellen; de gemiddelde effectgrootte van het gemiddeld prestatieverschil tussen PPT en CBT is meestal (verwaarloosbaar) klein en de spreiding van de effectgroottes is wisselend (in de ene review of meta-analyse wat groter dan in de andere);
- als er bij rekenen-wiskunde een significant moduseffect op de hoogte van de prestaties gevonden wordt, is dat veel vaker in het voordeel van PPT dan van CBT;
- het effect van de afnamemodus is sterk afhankelijk van het itemtype; er is nog weinig vergelijkingsonderzoek gedaan met andere vraagtypen dan meerkeuzevragen, maar de eerste resultaten lijken erop te wijzen dat moduseffecten bij meerkeuzevragen kleiner zijn dan bij andere itemtypen.

**Het toekennen van *partial credit* aan antwoorden**

De tweede onderzoeksvraag bestond uit twee delen:

- In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met het al dan niet toekennen van deelscores (*partial credit*) aan gedeeltelijk goede antwoorden?
- In hoeverre is het tegenwoordig mogelijk om de resultaten van de leerlingen op een CBT-examen op eenzelfde manier te beoordelen als de gebruikelijke *partial credit* beoordeling in een PPT-examen door beoordelaars.

De belangrijkste resultaten van de gevonden studies kunnen we als volgt samenvatten:

- Onder constant houding van de maximumscore (en het antwoordmodel) zou *partial credit* scoring waarbij gedeeltelijk goede antwoorden extra punten opleveren logischerwijs tot hogere scores moeten leiden dan alles-of-niets scoring van alleen het eindantwoord. Niet verrassend blijkt geautomatiseerde *partial credit* scoring te resulteren in hogere scores dan geautomatiseerde alles-of-niets scoring van alleen het eindantwoord. Dat het percentage goed op het examen door het toekennen van deelscores hoger wordt, betekent natuurlijk niet dat de kandidaten dan vaardiger zijn geworden. Het betekent wel dat de resultaten op het examen minder snel als tegenvallend geïnterpreteerd zullen worden.
- De laatste dertig jaar is er ervaring opgedaan met CBT waarbij optionele stapsgewijze bevraging wordt gecombineerd met geautomatiseerde scoring van gedeeltelijk goede antwoorden. Het blijkt mogelijk om de 'gewone' arbeidsintensieve PPT-beoordeling met deelscores in een automatisch gescoorde CBT met optionele *Steps* na te bootsen De resultaten wijzen erop dat deze nieuwe manier van examinering tot scores leidt die vergelijkbaar zijn met de gebruikelijke *partial credit* scoring van op papier gegeven antwoorden door menselijke beoordelaars.
- Stapsgewijze bevraging leidt tot een toename van de afnametijd.

**Mogelijkheden van item review**

In hoeverre hangen eventuele prestatieverschillen tussen PPT en CBT samen met de verschillende mogelijkheden tot item review? Onder item review wordt verstaan a) het overslaan van vragen om deze later te beantwoorden, b) het bekijken van eerder gegeven antwoorden en c) het veranderen van eerder gegeven antwoorden (Wise & Plake, 1989). Anders dan PPT biedt de eerste-generatie CBT geen mogelijkheden tot item review. De belangrijkste conclusies van het gevonden onderzoek kunnen als volgt worden samengevat:

- onderzoek met papieren toetsen laat zien dat veel leerlingen gebruik maken van item review, maar dat slechts een klein deel van de antwoorden daadwerkelijk gewijzigd wordt; en als er een antwoord gewijzigd wordt, is dat veel vaker van fout naar goed dan van goed naar fout; item review heeft hiermee een (klein) positief effect op de prestaties;
- er is onvoldoende evidentie gevonden om de vraag naar het eventueel verschillend gebruik van item review in PPT en moderne CBT met mogelijkheden van itemreview eenduidig te kunnen beantwoorden;
- als reguliere PPT en modernere CBT vergelijkbare mogelijkheden tot item review bieden, leidt dat doorgaans tot vergelijkbare resultaten;
- item review leidt tot een aanzienlijke toename van de afnametijd.

**Gebruik van kladpapier**

De vierde onderzoeksvraag betrof het eventueel verschillend gebruik van kladpapier in PPT en CBT en het differentiële effect daarvan op de prestaties. De veronderstelling bij CBT minder kandidaten van kladpapier gebruik maken dan bij PPT. Hiermee samenhangend zouden kandidaten bij CBT vaker van mentale oplossingsstrategieën ('uit het hoofd') gebruiken ten koste van schriftelijke oplossingsstrategieën ('op papier'). Onderzoek suggereert dat mentale oplossingsstrategieën doorgaans minder effectief zijn dan schriftelijke oplossingsstrategieën. De resultaten van een nog zeer beperkt aantal gevonden studies naar het mogelijk verschillend gebruik van kladpapier in PPT en CBT en het effect daarvan op de prestaties kunnen als volgt worden samengevat:

- wat betreft de vraag of kandidaten bij CBT meer of minder gebruik maken van kladpapier dan bij PPT geeft het gevonden onderzoek tegenstrijdige resultaten te zien;
- echter, als gevonden wordt dat kandidaten bij CBT minder gebruik maken van kladpapier dan bij PPT, heeft dat een hooguit zwak negatief effect op de prestaties; bij CBT lijken kandidaten dan wat vaker minder effectieve mentale oplossingsstrategieën te gebruiken en wat minder vaak schriftelijke oplossingsstrategieën;
- eventuele negatieve effecten van het verschillend gebruik van kladpapier zijn mogelijk afhankelijk van de specifieke rekenvaardigheid (bij bewerkingen en bij geometrische relaties en ruimtelijk redeneren groter dan bij andere rekenvaardigheden) en de sekse van de leerling (bij jongens groter dan bij meisjes).

Deze conclusies omtrent het gebruik van kladpapier moeten als zeer voorlopig worden beschouwd. Enerzijds omdat de literatuur search in dit geval verre van systematisch was waardoor het waarschijnlijk is dat relevante studies over het hoofd zijn gezien. Anderzijds omdat er nog maar weinig onderzoek lijkt te zijn gedaan waarin het verschillend gebruik van kladpapier op systematische wijze in verband wordt gebracht met prestatieverschillen tussen PPT en CBT. Zeldzaam lijken met name studies waarin het differentieel effect van kladpapier op de prestaties nauwkeurig gekwantificeerd wordt. Nog zeldzamer lijken methodologisch zuivere studies waarin onderzoeksontwerp en instrumentatie het toelaten om het effect van kladpapier systematisch te onderscheiden van andere verklaringen voor gevonden prestatieverschillen tussen PPT en CBT.

**Invoer van antwoorden via beeldscherm versus papier**

De vijfde onderzoeksvraag betrof de vraag in hoeverre prestatieverschillen op PPT en CBT samenhangen met de manier waarop kandidaten hun antwoord moeten geven: via het beeldscherm versus op papier. Kandidaten hebben doorgaans meer ervaring met het noteren van antwoorden op papier dan met het invoeren van antwoorden via het beeldscherm. De computer zou ten onrechte antwoorden fout rekenen die een menselijke beoordelaar op papier (gedeeltelijk) goed zou hebben gerekend, met als gevolg lagere prestaties voor CBT in vergelijking met PPT. De belangrijkste, nog zeer voorlopige conclusies van een nog zeer klein aantal gevonden studies vatten we als volgt samen:

- er lijken zwakke aanwijzingen te zijn dat geautomatiseerde scoringsalgoritmen antwoorden (wel eens) fout rekenen die menselijke beoordelaars op papier (gedeeltelijk) goed zouden rekenen;
- er zijn echter te weinig studies gevonden om met enige zekerheid uitspraken te kunnen doen over de mate waarin dit voorkomt en hoe groot het eventuele effect ervan is op de prestaties;
- er zijn *input tools* en feedback-mechanismen ontwikkeld die mismatch tussen computer en kandidaat minimaliseren en die kandidaten de mogelijkheid bieden om kleine invoerfoutjes met aftrek van punten te herstellen; effectiviteit en eventuele neveneffecten (bijvoorbeeld gokgedrag) zijn evenwel ongewis.

Deze conclusies betreffende miscommunicatie tussen computer en kandidaat bij de invoer van antwoorden en het effect ervan op de prestaties moeten als zeer voorlopig worden beschouwd. De literatuur search was te weinig systematisch en het leidt geen twijfel dat relevante studies zijn gemist.

# 5 Geraadpleegde literatuur


Abels, M., Boon, P., & Tacoma, S. (2013). *Basishandleiding Digitale Wiskunde Omgeving*. Utrecht: Freudenthal Instituut.

AERA American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA

APA American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: APA.

Ashton, H.S., Beevers, C.E., Korabinski, A.A. & Youngson, M.A. (2006). Incorporating partial credit in computer-aided assessment of mathematics in secondary education. *British Journal of Educational Technology*, *37*, 1, 93-119.

Béguin, A.A., & Wools, S. (2015). Vertical comparison using reference sets. In: R.E. Millsap, L.A. van der Ark, D.M. Bolt, & W.C. Wang (Eds.), *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics* (pp. 195-211). New York: Springer.

Benjamin, L.T., Cavell, T.A., & Shallenburger, W.R. I. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 3, 133-141.

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6,* 9.
Beschikbaar via http://www.jtla.org.

Bergstrom, B.A., & Lunz, M.E. (1992). Confidence in pass/faildecisions for computer adaptive and paper and pencil examinations. *Evaluation and the Health Professions,15*, 4, 453-464.

Braswell, J.S, & Bridgeman, B. (1992). *Effects of scratchwork space on SAT-M performance (ETS-report nr RM-92-10)*. Princeton: ETS.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.), *Educational measurement, Third Edition (*pp. 367-407). London: Collier Macmillan.

Cito (2015a). *Resultaten van de vragenlijst voor leerlingen die deelnamen aan de Rekentoets Voortgezet Onderwijs maart 2015*. Arnhem: Cito.

Cito (2015b). *Antwoordenanalyse rekentoets 2F voortgezet onderwijs. Weergave en analyse van antwoorden op de voorbeeldtoets 2F 2014*. Arnhem: Cito.

Csapó, B., Molnár, G., & Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: a pilot study for introducing TAO in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: new approaches to skills assessment and implications for large-scale testing* (pp. 120-126) Luxembourg: Office for Official Publications of the European Community. Beschikbaar via http://www.gesci.org/assets/files/reporttransition.pdf

College voor Examens (2014). *Tussenrapportage centraal ontwikkelde examens mbo en Rekentoets VO, 2013-2014*. Utrecht: College voor Examens.

College voor Toetsen en Examens (2014). *Rapportage referentiesets taal (lezen) en rekenen*. Utrecht: CvTE.

Commissie Bosker (2014). *Advies over de uitwerking van de referentieniveaus 2F en 3F voor rekenen in toetsen en examens*. Enschede: SLO.

Darrah, M., Fuller, E., & Miller, D. (2010). A comparative study of partial credit assessment and computer-based testing for mathematics. *Journal of Computers in Mathematics and Science Teaching, 29*, 4, 373-398.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*, 1297–1326.

Dimock, P.H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement & Evaluation in Counseling & Development, 24*, 119–126.

Eaves, R. C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. *Journal of Experimental Education, 55*, 23–26.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)*. Amsterdam: Faculteit der Maatschappij- en Gedragswetenschappen.

Fiddes, D.J., Korabinski, A.A., McGuire, G.R., Youngson, M.A., & McMillan, D. (2002). Does the mode of delivery affect mathematics examination results? *Alt-J, 10, 1*, 60-69.

Gallagher, A., Bridgeman, B., & Calahan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender and language groups* (RR–00–8). Princeton, NJ: Educational Testing Service.

Gaskill, J., & Marshall, M. (2006). *Comparisons between paper- and computer-based tests: A Literature Review.* Kelowna, BC, Canada: Society for the Advancement of Excellence in Education.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L, & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-359.

Greenwood, L., McBride, F., Morrison, H., Cowan, P. & Lee, M. (2000). Can the same results be obtained using computer-mediated tests as for paper-based tests for National Curriculum Assessment? In *Proceedings of International Conference on Mathematics / Science Education and Technology 2000* (pp. 179-184). Association for the Advancement of Computing in Education (AACE).

Greaud, V., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23–34.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research, 46*, 29-42

Harvey, A. L. (1987). *Differences in response behavior for high and low scores as a function of item presentation on a computer assisted test*. Unpublished doctoral dissertation. Lincoln: University of Nebraska.

Hamilton, L.S., Klein, S.P., & Lorie, W. (2000). *Using web-based testing for large-scale assessment*. Santa Monica, CA: RAND Corporation.

Harmes, J.C., & Parshall, C.G. (2000, November). *An iterative process for computerized test development: Integrating usability methods*. Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee.

Hickendorff, M., Heiser, W.J., Van Putten, C.M., & Verhelst, N.D. (2009). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika, 74*, 2, 331-350.

Hickendorff, M., Putten, C.M. van, Verhelst , N.D., & Heiser, W.J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology, 102*, 2, 438-452.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826–838.

International Test Commission (2001). International guidelines for test use. *International Journal of Testing, 1,* 93–114.

International Test Commission (2006). International guidelines on computer-based and Internet delivered testing. *International Journal of Testing, 6*, 143–172.

Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4. Uitkomsten van de vierde peiling in 2004* (PPON-reeks nummer 32). Arnhem: Cito.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and*

*Assessment*, *4,* 5, 1-34.

Johnson, D.E., & Mihal, W.L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist, 28,* 8, 694–699.

Keng, L., McClarty, K.L., & Davis, L.L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education, 21,* 3, 207-226.

Kim, J. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Kingston (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*, 22-37.

Kolen, M.J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment 6*, 73-96.

Lee, J. (1986). The effects of past computer experience on computer aptitude test performance. *Educational and Psychological Measurement, 46*, 727–733.

Lee, J. A., & Hopkins, L. (1985). *The effects of training on computerized aptitude test performance and anxiety.* Paper presented at the annual meeting of the Eastern Psychological Association (Boston, MA, March).

Lee, J.A., Moreno, K.E, & Sympson, J.B. (1986). The effects of test administration on test performance. *Educational and Psychological Measurement, 46,* 2, 467-474.

Leeson, H.V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6,* 1, 1-24.

Luecht, R.M., Hadadi, A., Swanson, D. B., & Case, S.M. (1998). Testing the test: A comparative study of a comprehensive basic science test using paper-and-pencil and computerised formats. *Academic Medicine, 73,* 51–53.

Mason, B.J., Patry, M., & Bernstein, D.J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24,* 1, 29-39.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McGuire, G.R., Youngson, M.A., Korabinski, A.A., & McMillan, D. (2002). Partial credit in mathematics exams: a comparison of traditional and CAA exams, *Proceedings of the 6th International CAA Conference*, Loughborough University.

Mueller, D.J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14,* 1, 9–14.

Murphy, P.K., Long, J., Holleran, T., & Esterly, E. (2000, August*). Persuasion online or on paper: A new take on an old issue*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Olea, J., Revuelta, J., Ximénez, M.C., & Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica,* 21, 157-173.

O'Malley, K.J., Kirkpatrick, R., Sherwood, W., Burdick, H.J., Hsieh, M.C., & Sanford, E.E. (2005, April). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Oregon Department of Education (2007). *Comparability of student scores obtained from paper and computer* administrations. Salem, OR: Oregon Department of Education.
Beschikbaar via
http://www.ode.state.or.us/teachlearn/testing/manuals/2007/doc4.1comparabilitytesatopandp.pdf

Pass-it (2002). *Good practice guide in question and test design.* Luton: CAA Centre.
http://www.calm.hw.ac.uk/GeneralAuthoring/031112-goodpracticeguide-hw.pdf

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Passmore, T., Brookshaw, L, & Butler, H. (2011). A flexible, extensible online testing system for mathematics. *Australasian Journal of Educational Technology, 27*, 6, 896-906.

Paek, P. (2005). *Recent trends in comparability studies.* Austin, TX: Educational Measurement.

Pearson (2009). Computer-based & paper-pencil test comparability studies. *Test, Measurement & Research Services Bulletin, 9*.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment 3, 6*.
Beschikbaar via http://www.jtla.org

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment 2,* 6, 1-45.

Putten, C.M. van (2008). De onmiskenbare daling van het prestatiepeil bij de bewerkingen sinds 1987 – een reactie. *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk, 27,* 1*,* 35 – 40.

Revuelta, J., Ximénez, M.C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement, 63*, 791–808.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education, 10, 3,* 279–294.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005-457). Washington, DC: Department of Education, National Center for Education Statistics.

Sangwin, C.J., & Ramsden, P. (2007). Linear syntax for communicating elementary mathematics. *Journal of Symbolic Computation, 42*, 9, 920-934.
Scheltens, F., Hickendorff, Eggen, Th. & Hiddink, L. (2014). Hoofdrekenen met papier - hoe zit dat met leerlingen die scoreenen? *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk, 33*, 128-140.

Scheltinga, F., Keuning, J., & Kuhlemeier, H. (2014). *Gericht werken aan opbrengsten in taal- en leesonderwijs: Een systematische review naar toetsvormen*. Cito/Expertisecentrum Nederlands: Arnhem/Nijmegen.

Schwartz, S.P., McMorris, R.F., & DeMers, L.P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement*, *28*, 2, 163-171.

Sutherland-Smith, W. (2002). Weaving the literacy web: changes in reading from page to screen. *The Reading Teacher, 55, 7,* 664–667.

Spray, J.A., Ackerman, T.A., Reckase, M.D., & Carlson, J.E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261–271.

Straat, H. (@in voorbereiding@). @@@. Arnhem: Cito.

Threfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66, 335-348.*

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In Bennett, R., & Ward, W. (eds.). *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van den Heuvel-Panhuizen, M., & Bodin-Baarends C. (2004). All or nothing: Problem solving by high achievers in mathematics. *Journal of the Korea Society of Mathematical Education, 8,* 3, 115-121.

Vispoel, W.P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35*, 328–345.

Vispoel, W.P., Wang, T., De la Torre, R., Bleiler, T., & Dings, J. (1992). *How review options and administration modes influence scores on computerized vocabulary tests*. Paper presented at the annual meeting of the National Council on Measurement in Education (San Francisco, CA).

Vispoel, W.P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60*, 371–384.

Way, W. D., Davis, L.L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA).

Wild, D.G., Beevers, C.E., Fiddes, D.I., McGuire, G.R. and Youngson, M.A. (1997). *Interactive PastPapers for A-Level and Higher Mathematics.* Glasgow: Lander Educational Software.

Wise, S.L., & Plake, B.S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, *8*, 3, 5-10.