☑ SAMENVATTING (NEDERLANDS)

Feedback is de krachtigste motor van elk leerproces. In de wiskundedidactiek wordt daarom uitgebreid onderzocht hoe men beoordelingen kan automatiseren. Dat is niet evident voor leerlingen: zich wiskundig uitdrukken is moeilijk op een computer en leersystemen kunnen vaak enkel de uitkomst verwerken en niet de oplossingsmethode. Digitale testen blijken zich veelal te beperken tot procedurele kennis ten koste van inzichtelijke denkvragen. Digitale wiskundetests ontwikkelen is een tijdrovende klus en daarnaast zijn leraars erg sceptisch om ze in te zetten, waardoor pen-en-papier het wiskundeonderwijs nog steeds domineert. Eén van de karakteristieken van wiskundig beoordelingswerk is dat foute antwoorden in een klasgroep patronen vertonen. Bijgevolg moeten leerkrachten hun feedback en punten meermaals herhalen. Dat brengt ons op het idee van semi-automatisch beoordelen: door handgeschreven leerlingenoplossingen digitaal te beoordelen, kan feedback bewaard en vervolgens hergebruikt worden. Dat kan uitgebreidere feedback, tijdswinst en verhoogde interbeoordelaarsbetrouwbaarheid opleveren. In dit proefschrift werden twee semi-automatische beoordelingsmethoden ontwikkeld en onderzocht.

Voor de eerste studie werd een softwaretool geprogrammeerd die de feedback van de leerkrachten tijdens het nakijken opslaat, zodat die makkelijk hergebruikt kan worden als dezelfde of gelijkaardige fouten zich opnieuw aandienen. Om leerkrachten te leren hoe zij herbruikbare feedback kunnen opstellen, werd atomische feedback uitgevonden. Om feedback atomisch te schrijven, moeten de leraars de verschilllende fouten van een student identificeren en korte feedbackitems schrijven voor elke afzonderlijke fout. Hierbij is het van belang dat deze dat deze items onafhankelijk van elkaar zijn. Items die thematisch of in de oplossing van de leerling bij elkaar horen, kunnen worden geclusterd tot een hiërarchische lijst. Tijdens het onderzoek konden we aantonen dat atomische feedback de herbruikbaarheid van feedback significant verbetert. Bovendien konden leraars zich deze vormvereisten snel eigen maken: na slechts een korte introductie, kon tijdens een cross-overexperiment met 45 wiskundeleerkrachten, 74% van de gegeven feedback met de feedbacktool als atomisch geclassificeerd worden. De 26% niet-atomische items behandelden vaak meerdere fouten binnen één item of vermeldden zowel de fout als de plaats ervan in de oplossing van de leerling. In de geest van atomische feedback, moeten deze twee verdeeld worden over twee feedbackitems: één voor de locatie en één voor de fout. Ze kunnen in de lijst van feedbackitems aan elkaar worden gelinkt door het item over de fout te laten inspringen onder het item met de locatie.

Tijdens het experiment werd een opvallende ontdekking gedaan: de semi-automatische feedbacktool leidde ertoe dat leerkrachten aanzienlijk meer feedback gaven in plaats van tijd te besparen in vergelijking met het opstellen van klassieke, handgeschreven feedback. Deze twee feedbacktypes werden ook vergeleken naar vorm en inhoud met behulp van text mining en kwalitatieve technieken. Woordfrequenties, gevoelens en de hoeveelheid foutieve, beschrijvende en corrigerende feedback waren vergelijkbaar in beide feedbacktypes. Wanneer leerkrachten de semi-automatische tool gebruikten, was de feedback uitgebreider maar minder specifiek gericht op de oplossing van de leerling. Zonder de tool was de feedback korter, concreter en meer gefocust op de hoofdzaken. De eigenschappen van de feedback met de semi-automatische tool waren zeker niet altijd beter dan die van de klassieke feedback: de leerkrachten hadden vaker de neiging het werk van de leerlingen te beschrijven en te corrigeren in plaats van de onderliggende misconcepties te analyseren. In het algemeen mogen leerkrachten de handigheid van de tool niet verwarren met kwaliteit: een grote feedbackvaardigheid, vakinhoudelijke en vakdidactische kennis blijven essentiële leerkrachtkenmerken om kwaliteitsvolle feedback te garanderen.

De tweede studie was een samenwerking met de Examencommissie Secundair Onderwijs van de Vlaamse overheid. Tijdens het onderzoek werd hun traditionele beoordelingsmethode voor handgeschreven wiskunde-examens omgevormd tot een semi-geautomatiseerd systeem die omgedoopt werd tot 'checkbox grading.' Elke corrector ontvangt een lijst met aankruisvakjes en moet diegene aanvinken die van toepassing zijn op de leerlingenoplossing. Er kunnen afhankelijkheden tussen deze vakjes worden gedefinieerd om ervoor te zorgen dat alle beoordelaars dezelfde weg afleggen doorheen het beoordelingsschema. Het systeem berekent automatisch het cijfer en genereert feedback die een gedetailleerd inzicht geeft in wat er fout ging en hoe het cijfer werd beoaald, gebaseerd op vooraf opgestelde atomische feedback. Bij de traditionele beoordelingsmethode delen de beoordelaars alleen een cijfer mee op basis van beoordelingsschema's die door de examencommissie werden opgesteld. De methode werd onderzocht zowel vanuit het oogpunt van de correctoren, als vanuit het oogpunt van de leerlingen (kandidaten).

We onderzochten de tijdsbesteding, waardering en interbeoordelaarsbetrouwbaarheid van de correctoren. Het nakijken met 'checkbox grading' duurde ongeveer twee keer zo lang als het beoordelen op de traditionele manier. Verrassend genoeg was het subjectieve tijdsbesef van de correctoren in tegenspraak met deze metingen: zij rapporteerden net dat ze sneller hun taken als corrector konden uitvoeren met 'checkbox grading'. Het is mogelijk dat de correctoren de transparantie van het beoordelingswerk en de daaruit voortvloeiende feedback voor de kandidaten van hoger belang achtten dan de tijd die ze daarvoor nodig hadden. Dit blijkt ook uit hun algemeen hoge waardering voor de semi-automatische beoordelingsmethode.

De interbeoordelaarsbetrouwbaarheid van de correctoren werd vergeleken tussen blind versus zichtbaar beoordelen met 'checkox grading' enerzijds, en met de traditionele beoordelingsmethode anderzijds. Bij het beoordelen met 'checkox grading' kunnen de aankruisvakjes gekoppeld worden aan deelcijfers. Door de aankruisvakjes aan te klikken die op de oplossing van de leerling van toepassing zijn, berekent de computer automatisch het totaalcijfer. Dit leidde tot het experimentele idee van blind beoordelen waarbij noch de deelcijfers, noch het totaalcijfer aan de correctoren werden getoond. Uit de literatuur over rubrics is bekend dat beoordelaars durven afwijken van de criteria

wanneer hun holistische appreciatie van het werk van een leerling niet in overeenstemming is met de evaluatie die uit de rubric voortvloeit. Door de cijfers te verbergen, wilden we onderzoeken of we dit cognitieve conflict konden vermijden, wat zou moeten resulteren in een hogere interbeoordelaarsbetrouwbaarheid. Er bleek echter geen geschikte interbeoordelaarsbetrouwbaarheidsmaat te bestaan om deze onderzoeksvraag te beantwoorden. Er bestond namelijk geen voor toeval gecorrigeerde κ-coëfficiënt waarmee de betrouwbaarheid van meerdere beoordelaars, die voor elke leerlingenoplossing één of meerdere aankruisvakjes selecteren, kon worden berekend. Bekende maten zoals de Cohen's kappa en Fleiss' kappa laten slechts de selectie van één item per leerlingenoplossing toe. Dit leidde tot de ontdekking van een gegeneraliseerde Fleiss' kappa, die rekening houdt met alle informatie die 'checkbox grading' oplevert. Deze maat stelde ons in staat de onderzoeksvraag te beantwoorden: blind beoordelen verbetert de interbeoordelaarsbetrouwbaarheid wanneer beoordelingsschema's streng zijn en strikt moeten worden geïnterpreteerd, terwijl zichtbaar beoordelen beter geschikt is voor complexere beoordelingsschema's, omdat het zien van de cijfers de beoordelaars helpt de juistheid van hun eigen beoordelingswerk in te schatten. Vergeleken met de traditionele methode was 'checkbox grading' even betrouwbaar.

Om te onderzoeken hoe leerlingen reageren op de resulterende atomische feedback van 'checkbox grading', werd een vragenlijst afgenomen bij 36 van de 60 leerlingen die deelnamen aan het examen waarop dit onderzoek betrekking had. Vier van hen stemden in met semi-gestructureerde interviews. Leerlingen gaven de voorkeur aan de traditionele beoordelingsschema's boven 'checkbox grading' wanneer hen werd gevraagd feedbacksoorten te rangschikken van meer naar minder begrijpelijk. Toen leerlingen echter werden geïnterviewd met behulp van een think-aloud protocol, bleek dat ze 'checkbox grading' feedback gemakkelijker konden interpreteren. Bovendien waren 97% van de leerlingen het er op de vragenlijst mee eens dat de Examencommissie 'checkbox grading' zou moeten gebruiken als standaard beoordelingsmethode. Vooral de duidelijke link tussen de feedback en de totstandkoming van het cijfer werd hoog ingeschat. Hun begrip van dit soort feedback was gemiddeld hoog en kon niet gecorreleerd worden met hun examencijfer. Dit betekent dat nagenoeg alle leerlingen, ook de minder goed presterende, de resulterende feedback vlot geïnterpreteerd krijgen.

De twee semi-automatische beoordelingsmethoden tonen twee waardevolle manieren waarop computers en leerkrachten kunnen samenwerken bij het beoordelen en geven van feedback aan leerlingen in het wiskundeonderwijs. De eerste aanpak was gebouwd voor individuele leerkrachten, de tweede voor een groep correctoren. Uit beide studies blijkt dat het geven van feedback altijd meer werk vereist dan het louter aanduiden van fouten of het meedelen van een cijfer. Bovendien lijken semi-automatische hulpmiddelen leraren vaak te motiveren om nog meer werk te doen in plaats van tijd te besparen. We concluderen met nog enkele ideeën voor vervolgonderzoek. Een eerste prioriteit is het slimmer maken van het suggestiesysteem van de eerste semi-geautomatiseerde tool. Door de ideeën uit de literatuur rond aanbevelingssystemen te integreren, kan het selecteren van feedback items om te hergebruiken vlotter verlopen. Daarnaast kunnen semi-automatische beoordelingsmethoden gelinkt worden aan Bayesiaanse netwerken om scherper zicht te krijgen op het individuele leerproces van leerlingen. Een Bayesiaans netwerk is een probabilistisch, grafisch model dat de bekwaamheid van een leerling in kaart brengt. Ook het onderzoeken van semi-geautomatiseerd beoordelen in andere settings (zoals peer feedback) zijn vruchtbare grond voor vervolgonderzoek.