10

Het meten van veranderingen

In het onderwijs kan een groeiende belangstelling bespeurd worden voor systemen die de vorderingen van individuele leerlingen kunnen meten. Zulke systemen noemt men leerlingvolgsystemen (LVS). Daarbij gaat het om de volgende vragen. Hoeveel beter kan een leerling technisch lezen na drie maanden onderwijs? In welke mate is de leerling het afgelopen half jaar vooruitgegaan in rekenen? Deze vragen refereren aan veranderingen in individuele vaardigheidsniveaus. We proberen dan individuele groei, op basis van meetresultaten op verschillende tijdstippen, te kwantificeren. In het verleden was de gangbare praktijk groei te meten met veranderingsscores, het verschil tussen twee meetresultaten, meestal binnen het kader van de klassieke testtheorie. Het meten van groei met veranderingsscores was echter geen succes. Vandaar dat wij in dit hoofdstuk een meer modelmatige benadering kiezen, veranderingsscores blijven buiten beschouwing.

We gaan na wat de meetmodellen die in de hoofdstukken 3 en 4 zijn besproken, de klassieke testtheorie en de itemresponstheorie te bieden hebben voor het volgen van individuele vaardigheden. In principe zijn deze meetmodellen statisch, dat wil zeggen: ontworpen voor metingen op één bepaald tijdstip. Een meetmodel beschrijft de relatie tussen het meetresultaat en de te meten vaardigheid op één tijdstip, bijvoorbeeld de relatie tussen observatie en ware score (klassieke testtheorie) of latente vaardigheid (itemresponstheorie). Bij het meten van veranderingen beschikken we over meetresultaten van hetzelfde individu op verschillende tijdstippen. Toepassing van een statisch meetmodel op de meetresultaten resulteert dan in een aantal momentopnamen van de te meten vaardigheid, zonder er rekening mee te houden dat de metingen betrekking hebben op hetzelfde individu. Modellen die metingen aan hetzelfde individu op meer dan een tijdstip beschrijven, worden aangeduid als dynamische of tijdsafhankelijke modellen. Dynamische modellen onderscheiden zich van statische modellen door expliciet de relatie te leggen tussen metingen op verschillende tijdstippen.

In dit hoofdstuk ligt de nadruk op modellen die de vorderingen in leerresultaten van individuele leerlingen kunnen beschrijven of voorspellen. In de eerste paragraaf wordt

de problematiek van het meten van veranderingen in het algemeen besproken. De bepaling van individuele vorderingen wordt, met als uitgangspunt een simpel lineair groeimodel, in de tweede paragraaf uitgewerkt, waarbij als meetmodel de klassieke testtheorie wordt gehanteerd. Hetzelfde doen we in de derde paragraaf, maar nu met een itemresponsmodel als meetmodel. Het accent in de paragrafen 10.2 en 10.3 ligt op de vergelijking van een statische en een dynamische aanpak bij de modellering en de consequenties daarvan voor de bepaling van individuele vorderingen. Tenslotte wordt in de laatste paragraaf de problematiek van het meten van veranderingen in een breder perspectief geplaatst en wordt nader ingegaan op alternatieve benaderingen en verwachtingen over mogelijke ontwikkelingen.

10.1 Individuele groei

De problematiek van het meten van veranderingen, het volgen van leerresultaten, of meer algemeen het vaststellen van groei, is geen sinecure. In het verleden zijn sommige auteurs (Cronbach & Furby, 1970) zo pessimistisch geworden dat zij hebben voorgesteld de hele kwestie van veranderingsscores maar te vergeten en de onderzoeksvragen zo te formuleren dat er geen veranderingsscores aan te pas komen (zie ook Jansen, 1979). Uit het aantal verwijzingen naar het werk van Cronbach en Furby in recenter literatuur blijkt echter dat door de jaren heen de kwestie van het meten van veranderingen de wetenschap is blijven boeien.

In deze paragraaf onderzoeken we waar de problemen zitten bij het meten van verande- ringen. Eerst kijken we naar de relatie tussen model en data in een longitudinaal onderzoek. Daarna worden aan de hand van een concreet voorbeeld enkele problemen bij het meten van veranderingen geïllustreerd. De paragraaf wordt besloten met een korte verhandeling over de methodologische aspecten bij het meten van veranderingen, maar dan specifieker gericht op het volgen van individuele leerresultaten.

10.1.1 Longitudinale data en modellering

Als over een longitudinale gegevensverzameling wordt gesproken, wordt daarmee bedoeld dat men beschikt over meetresultaten van hetzelfde object met betrekking tot een bepaald attribuut op verschillende tijdstippen. In het onderwijs resulteert dit

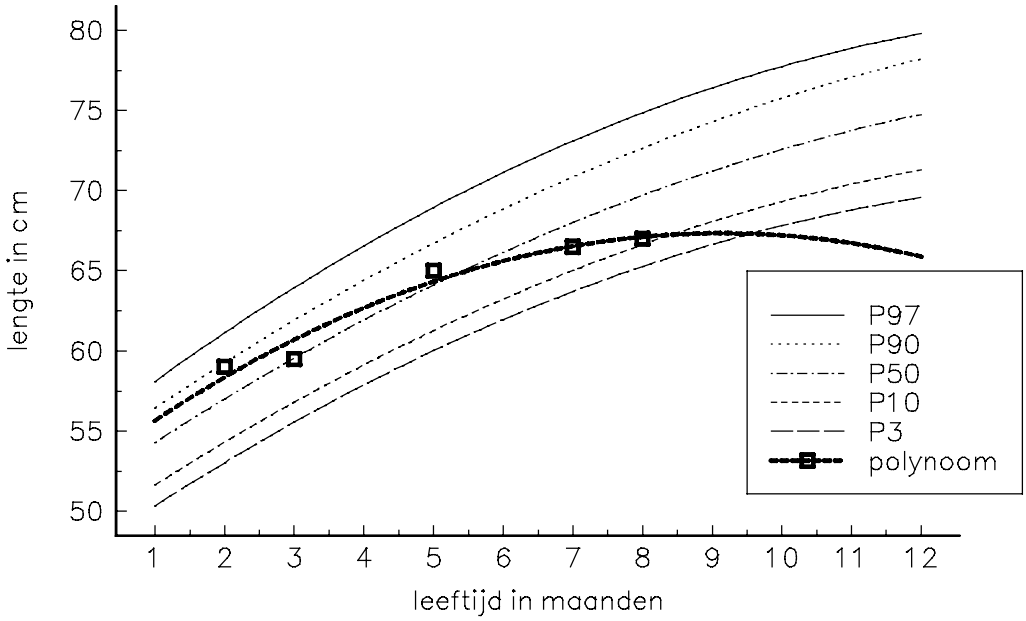
meestal in een gegevensver- zameling die betrekking heeft op de interactie tussen toetsen en individuen op verschillende tijdstippen.

Als men beschikt over een longitudinale dataset, is dat geen garantie dat die gegevens daad- werkelijk dynamisch gemodelleerd worden, dat wil zeggen dat de interactie tussen toetsen, individuen en tijdstippen simultaan wordt beschouwd. De gangbare praktijk is om longi- tudinale meetresultaten te reduceren tot statische momentopnamen op afzonderlijke tijdstippen. Eigenlijk beschouwen we dan de afzonderlijke meetmomenten los van elkaar. De afzonderlijke meetmomenten in de longitudinale gegevensverzameling duidt men aan met de term cross-secties. Wordt er al gebruik gemaakt van een dynamisch model, dan heeft dit vaak alleen betrekking op geaggregeerde gegevens op populatieniveau. Een voorbeeld van zo'n gangbare praktijk is het verticaal equivaleren (zie hoofdstuk 8). Binnen de itemresponstheorie is het mogelijk, onder zekere condities, een longitudinale dataset met een statisch meetmodel te analyseren. Feitelijk wordt de longitudinale gegevensverzameling opgedeeld in afzonderlijke cross-secties (individuen × toetsen) welke vervolgens worden gecombi- neerd in een onvolledig design tot één dataverzameling, die traditioneel met een statisch meetmodel geanalyseerd kan worden (zie bijvoorbeeld figuur 8.8 in hoofdstuk 8). Deze benadering is prima zolang zij schattingen van itemparameters en dergelijke betreft en we ons realiseren dat deze schattingen betrekking hebben op de onderhavige populatie. Bovendien geldt dat bij het analyseren van cross-secties van de data met een statisch model een mogelijke samenhang tussen de individuele meetresultaten in de tijd niet naar voren gehaald en belicht wordt. Veel van de door Cronbach en Furby (1970) gesignaleerde problemen bij het meten van veranderingen zijn dan ook artefacten van de gekozen benadering. Concluderend kan gezegd worden dat longitudinale gegevens in principe om een dynamisch model vragen.

10.1.2 Het vaststellen van de individuele groei bij zuigelingen

Op het consultatiebureau van de Kruisvereniging houdt men periodiek, naast andere zaken als gewicht en hoofdomtrek, de lichaamslengte van zuigelingen bij. Het doel hiervan is het tijdig signaleren van stagnaties in de groei zodat, indien gewenst, passende maatregelen genomen kunnen worden. De vraag rijst dan uiteraard wanneer er actie ondernomen dient te worden. We zullen hier niet de medische aspecten doch de methodologische aspecten beschouwen. De zuigeling wordt bij alle gelegenheden gemeten met dezelfde schuifmaat met een schaal in centimeters. Laten we aannemen dat bij de metingen de meetfout verwaarloosbaar is. Het is duidelijk dat bij alle

gelegenheden een en het zelfde attribuut, lichaamslengte in cm, bij de zuigeling gemeten wordt. In figuur 10.1 hebben we voor een hypothetische zuigeling de gemeten lichaamslengte uitgezet tegen de leeftijd in maanden. De open vierkantjes zijn de waarnemingen bij de leeftijden: 2, 3, 5, 7 en 8 maanden. De mate van groei kunnen we direct aflezen als het verschil tussen twee metingen. Na drie maanden meet de zuigeling 59.5 cm en na vijf maanden 65 cm: in twee maanden tijd is de zuigeling 5.5 cm gegroeid. Zou de medicus over absolute criteria beschikken, bijvoorbeeld dat na zeven maanden elke zuigeling 60 cm moet zijn, dan is het mogelijk op grond hiervan te beslissen of voor een specifieke zuigeling hulp nodig is. Aangezien absolute criteria meestal niet voorhanden zijn, gebruikt men relatieve. Men zou bijvoorbeeld de populatie zuigelingen in Nederland kunnen beschouwen en met behulp van een steekproef kunnen vaststellen hoe de ontwikkeling in de populatie van zuigelingen er uit ziet. De ontwikkeling in de populatie kan men dan per tijdstip met referentiegegevens beschrijven, bijvoorbeeld door per tijdstip decielen of percentielen (zie paragraaf 13.4.1) te bepalen. Het signaleren van stagnatie in de groei kan dan relatief plaatsvinden, een afwijking van twee of meer decielen naar beneden zou men als ongewenst kunnen bestempelen. In figuur 10.1 zijn als referentiegegevens vijf percentiellijnen getrokken. De percentiellijn P50 bijvoorbeeld geeft aan waar het vijftigste percentiel voor een bepaalde leeftijd ligt. Met behulp van deze lijnen is het mogelijk de relatieve positie van de zuigeling aan te geven. In het voorbeeld bevindt de zuigeling zich na vijf maanden tussen de P50 en P90, na zeven maanden tussen de P10 en P50.



Figuur 10.1
Groeicurve voor een hypothetische zuigeling met referentiegegevens

De positie van de zuigeling in de Nederlandse populatie van zuigelingen is dus veranderd. Immers, na vijf maanden behoorde de zuigeling tot de 'groten', terwijl na zeven maanden de zuigeling bij de 'kleintjes' gerekend mag worden. Of deze ontwikkeling ongewenst is, is een medische vraag. Verder, maar meer discutabel op grond van het geringe aantal waarnemingen, is het mogelijk de groei van de zuigeling op de een of andere manier te modelleren. De meetpunten in figuur 10.1 zijn benaderd met een polynoom. Deze is zichtbaar als de dikke lijn. Het is nu mogelijk met behulp van dit polynoom, dat we kunnen opvatten als een groeimodel, predicties te doen. Op grond van dit simpele groeimodel is de verwachting dat de lichaamslengte van de onderhavige zuigeling na tien maanden ongeveer 67.5 centimeter is. Met behulp van predicties is het mogelijk reeds vooraf iets te signaleren: gegeven de curve tot nu verwachten we dat na tien maanden de zuigeling in de gevarenzone komt.

Er blijven nog genoeg vragen over. Bijvoorbeeld: is de Nederlandse populatie wel geschikt als referentiepunt? Denkbaar is dat een opdeling van de populatie naar geslacht of gewichts- klasse zeer zinvol zou kunnen zijn. Met andere woorden, niet één maar verschillende popu- laties worden beschouwd. Een complicerende factor in het voorbeeld is het feit dat groei bij de individuele zuigeling niet vloeiend, maar schoksgewijs verloopt. Voorstelbaar is dus dat ogenschijnlijke stagnatie, door het slecht kiezen van tijdstippen, ten onrechte tot de conclusie leidt, dat hulp geboden is. Iets

dergelijks zou men kunnen observeren in het voorbeeld: de lengte na twee en drie maanden is nagenoeg gelijk, terwijl we na vijf maanden een aanzienlijke groei zien.

Dit voorbeeld illustreert dat het vaststellen van (stagnaties in de) groei bij zuigelingen, ook al beschikken we over metingen met te verwaarlozen meetfouten, niet geheel vrij van proble- men is.

10.1.3 Problemen bij het volgen van individuele leerlingen

Waar gaat het nu precies om bij het volgen van de vaardigheid van individuele leerlingen? In eerste instantie proberen we de ontwikkeling van een vaardigheid, bijvoorbeeld het spellen van woorden, van een leerling in kaart te brengen. Afhankelijk van de resultaten kan men dan, net als in het voorbeeld bij de zuigeling, bepalen of deze ontwikkeling al dan niet voorspoedig verloopt en, zo nodig, proberen deze ontwikkeling bij te sturen. De ontwikkeling van de vaardigheid kan men opvatten als een gestructureerd proces waarvan de structuur nog gemodelleerd dient te worden. Modellen voor een gestructureerd proces worden aangeduid als groei-, proces-, tijdreeks- of structuurmodellen. In het onderwijs zal een groeimodel veelal op het niveau van de (sub)populatie geformuleerd zijn, daar we op het individuele niveau te weinig gegevens hebben om het proces te modelleren, dat wil zeggen een model te specifi- ceren, te schatten en te toetsen. Dit is het gevolg van het feit dat in het onderwijs het volgen van leerresultaten zich meestal beperkt tot twee à drie meetmomenten per jaar. Fraaier zou het zijn een leerling frequenter te toetsen. Het mag voor een ieder duidelijk zijn dat dit praktisch niet haalbaar en zelfs niet wenselijk is. In het meest extreme geval zou een leerling bij voortduring getoetst worden, van onderwijs zou dan geen sprake meer zijn. De dagelijkse evaluering van de ontwikkeling van de leerlingen moet hoe dan ook voorbehouden blijven aan de leerkracht. De consequentie hiervan is dat de toepassing van tijdreeksmodellen voor een individuele leerling niet mogelijk zal zijn. Immers, om tijdreeksmodellen zinvol te kunnen toepassen, moet de reeks een zekere minimale lengte hebben: bijvoorbeeld 50 waarnemingen. In het onderwijs, met twee à drie toetsmomenten per jaar, komen we vaak niet verder dan 10 à 15 waarnemingen per leerling gedurende de hele schooltijd. Als bij onderwijsdata de informatie voor een individuele leerling niet uit de lengte van de tijdreeks kan komen dan moet het maar uit de breedte komen! Gelukkig is dit mogelijk door individuele tijdreeksen te beschouwen als replicaties van een onderliggende tijdreeks op populatieniveau. Dit resulteert in een opzet met herhaalde metingen op het individuele niveau met replicaties op het niveau van de populatie.

In het voorbeeld van de lichaamslengte bij zuigelingen kan men de lengte direct waar- nemen. Bovendien kan de vergelijking van de lengte van twee zuigelingen zonder omweg plaatsvinden: leg ze naast elkaar. Om de groei van een zuigeling vast te stellen, een vergelijking van dezelfde zuigeling op twee tijdstippen, zullen we een meetinstrument moeten gebruiken. De keuze van een instrument om lengte te bepalen is niet problematisch. Voor de meting van lengte kunnen we terugvallen op internationaal gemaakte afspraken: lengte meten we in meters en de lengte van een meter ligt vast. Als de meeteenheid vastligt, resteert alleen nog de keuze van een adequaat meetinstrument. Dit meetinstrument moet geijkt zijn aan de standaardmeter, geschikt zijn voor de te meten objecten en zodanig zijn dat de afleesfout beperkt blijft. Voor de meting van lichaamslengte bij baby's kunnen we dan bijvoorbeeld een schuifmaat met een verdeling in centimeters nemen. Nu is het mogelijk de lichaamslengte van dezelfde baby in de tijd te vergelijken. In wezen zijn het meetprobleem, het nauwkeurig be- palen van de lengte op een tijdstip, en het groeiprobleem, de verandering van de lengte van een object tussen twee tijdstippen, gescheiden. Dit wil zeggen dat de meetfout die we maken geen systematische componenten bevat die afhankelijk zijn van het te meten object of de te meten grootheid.

De te modelleren processen in het onderwijs hebben meestal een latente structuur, daar de vaardigheden niet direct waarneembaar zijn. Bij latente vaardigheden als spellingvaardig- heid, zullen het meet- en het groeimodel in de regel niet gescheiden zijn. Allereerst dienen we indirect vast te stellen wat spellingvaardigheid is. Stel dat we beschikken over een valide meetinstrument, toets A, voor meetmoment 1. De vraag rijst hoe we kunnen weten of we op een later tijdstip nog dezelfde spellingvaardigheid meten als bij de eerdere afname. Afgezien van de vraag of we een leerling twee keer dezelfde toets kunnen voorleggen (denk bijvoorbeeld aan geheugeneffecten) is het evident dat we niet hetzelfde dictee kunnen afnemen bij groep 3 en groep 8. Een voor groep 3 geschikt dictee zal in groep 8 naar we hopen door een ieder foutloos gemaakt worden. Met andere woorden, we kunnen niet met één toets volstaan maar we zullen een hele batterij van toetsen moeten hebben. Problematisch is het nu deze toetsen aan elkaar te ijken. We beschikken namelijk niet, zoals bij de lengtemeting, over een standaardspellingvaardigheidsmeter. Het ijken van de toetsen zal nu expliciet in een meetmodel moeten gebeuren. Afhankelijk van het gekozen meetmodel en de daarin gehanteerde schattingsmethode, zal het niet altijd mogelijk zijn het meet- en het groeimodel gescheiden aan te pakken. Voordat we aan de modellering van groei toekomen, dienen er dus nog enkele problemen opgelost te worden met betrekking tot de validering en de ijking van de meetinstrumenten. In de eerste plaats: hoe kunnen we weten of we met verschillende toetsen dezelfde latente vaardigheid meten, zowel cross-

sectioneel als longitudinaal? En in de tweede plaats: hoe kunnen de behaalde resultaten bij die toetsen met elkaar vergeleken worden?

Een ander probleem bij de vaststelling van vorderingen in leerresultaten betreft in de termen van Bock (1976), de typische onbetrouwbaarheid van leerresultaten voor een individuele leerling. Als het gaat om groepsvergelijkingen of de normering van toetsen speelt deze onbetrouwbaarheid ons geen parten, maar op het individuele niveau des te meer. Als illustratie kan de standaardmeetfout in de klassieke testtheorie dienen. Bezien we de meet- resultaten van een leerling op twee tijdstippen en zetten we met behulp van de standaardmeetfout rond deze meetresultaten een betrouwbaarheidsinterval af, dan zien we dat deze inter- vallen elkaar zeer vaak overlappen, ook als het betrouwbare toetsen betreft. Statistisch gezien is er dan geen sprake van groei.

Gezien bovenstaande problemen zal het geen sinecure zijn om individuele groei vast te stellen. Om deze problemen te overwinnen is het nodig, zoals Bock al in 1976 constateerde, de aandacht in de psychometrie te verleggen. De aandacht zal verlegd moeten worden van statische momentopnames, de relatieve positie van leerlingen in een bepaalde groep, naar methoden en modellen die op adequate wijze de groei van individuele leerlingen kunnen beschrijven en voorspellen. Het gaat er om veranderingen in het traject dat een individuele leerling aflegt te detecteren.

Drie methodologische problemen bij het volgen van individuele leerlingen verdienen gerichte aandacht. In de eerste plaats is dat de formulering van adequate meetmodellen. Deze meetmodellen moeten in ieder geval informatie leveren over de precisie van een meetresultaat. Verder is het wenselijk dat de mate van precisie kan variëren over meetresultaten. Daarnaast moet het meetmodel de koppeling kunnen verzorgen tussen groeimodel en observaties. Een tweede aandachtspunt betreft de keuze van een geschikt groeimodel. Het is wenselijk dat het groeimodel flexibel is, in die zin dat groei voor individuen of groepen van individuen verschillend kan verlopen. Het derde aandachtspunt betreft de specificatie van wat in de literatuur een verfijnd referentiekader genoemd wordt. Hiermee bedoelen we dat het mogelijk moet zijn veranderingen in individuele groei af te zetten tegen relevante andere individuen, groepen en populaties en bovendien ook tegen nader te formuleren onderwijsinhoudelijke criteria.

In dit hoofdstuk zullen we het bepalen van individuele leerresultaten in de tijd uitwerken voor de twee meest gangbare meetmodellen in de psychometrie, te weten de klassieke testtheorie en de itemresponstheorie. We zullen daarbij rekening houden met de in deze paragraaf gesignaleerde problemen. Omwille van de eenvoud beperken we ons voor het groeimodel tot een lineair model voor één populatie. Verder blijven vragen aangaande validiteit nagenoeg buiten beschouwing, ervan uitgaande dat deze reeds elders beantwoord zijn.

10.2 Klassieke testtheorie en groeiscores

In deze paragraaf werken we de bepaling van groeiscores nader uit, waarbij we het model van de klassieke testtheorie als meetmodel hanteren. Aan de hand van gesimuleerde longitudinale data zal de schattingsproblematiek van de ware score doorlopen worden. Om voor deze data de groeiscores te bepalen worden twee benaderingen gebruikt: een statische en een dynamische. Recapitulerend luidt de vraagstelling: hoe schatten we de ware score als men de data behandelt als afzonderlijke momentopnamen en welke schatters komen voor de ware score in aanmerking als we de dynamiek in de data gebruiken?

10.2.1 Artificiële longitudinale data

Stel dat de heer Knikker over de uitzonderlijke gave beschikt om knikkervaardigheid bij kinderen direct en feilloos te kunnen vaststellen. Deze heer besluit te onderzoeken inhoeverre de psychometrici dat ook kunnen. Knikker is zich bewust van het unieke van zijn gave en begrijpt dat hij de psychometrici iets concreets in handen moet geven. Hij besluit daarom een experiment te doen. Op vier momenten in een leerjaar stelt hij bij een aselecte steekproef van 1000 kinderen uit groep drie van de basisschool de knikkervaardigheid vast. Deze ware knikkervaardigheidsscores houdt hij angstvallig geheim. Knikker is bekend met het feit dat psychometrici zich meestal met toetsscores moeten behelpen, daarom genereert hij op de vier momenten voor alle kinderen in de steekproef toetsscores volgens het klassieke meetmodel:

$$y_t = \eta_t + \varepsilon_t \qquad t = 1,2,3,4 \qquad \text{(meetvergelijking klassieke testtheorie)}$$

waarbij t het meetmoment aanduidt, y_t de toetsscore op meetmoment t, η_t de ware knikker- score op meetmoment t en ε_t een door de heer Knikker toegevoegde meetfout. Merk op dat wij hier voor een andere notatie van het klassieke meetmodel dan die in hoofdstuk 3 kiezen. Om verwarring te voorkomen tussen de in hoofdstuk 3 gebruikte letter T voor de ware score en de nu geïntroduceerde tijdstipindicator, t duiden we de ware score op tijdstip t in het vervolg aan met η_t. In tegenstelling tot hoofdstuk 3 worden de toetsscore X en de meetfout e nu aangeduid met respectievelijk y en ε. De gevolgde notatie is nu in overeenstemming met de gangbare notatie in lineaire structurele modellen (Jöreskog & Sörbom, 1989). De op deze manier gegenereerde toetsscores stelt Knikker beschikbaar. Om het de psychometrici

makkelijker te maken, laat hij weten dat de toetsscores zijn gegenereerd volgens bovenstaande meetvergelijking. Verder geeft hij aan dat de meetfouten onafhankelijk zijn van de knikkervaardigheidsscores, tussen meetmomenten ongecorreleerd zijn en bovendien normaal verdeeld zijn met verwachting 0 en gelijke variantie voor alle meetmomenten. Bovendien wordt de meetfoutvariantie gegeven, $\sigma_\varepsilon^2 = 6.25$. Verder wordt ook nog bekend gemaakt dat de ware knikkervaardigheid $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)'$, multivariaat normaal $N(\boldsymbol{\mu}_\eta, \Sigma_\eta)$ verdeeld is met

$$\boldsymbol{\mu}_\eta = \begin{pmatrix} 20 \\ 30 \\ 40 \\ 50 \end{pmatrix} \quad \text{en} \quad \Sigma_\eta = \begin{pmatrix} 25 & & & \\ 20 & 25 & & \\ 16 & 20 & 25 & \\ 12.8 & 16 & 20 & 25 \end{pmatrix}.$$

De vraag die de heer Knikker de psychometrici voorlegt is nu: wat zijn de ware knikkervaardigheidscores van deze kinderen op de vier meetmomenten? Twee teams van psychometrici, team A en team B, buigen zich over het probleem. Hierbij hanteert team A een statische benadering en team B een dynamische benadering. We zullen zien waarin het een en ander resulteert.

10.2.2 Statische benadering

De benadering van het probleem door team A is als volgt: men beschouwt de toetsscores op de afzonderlijke momenten als cross-secties. De longitudinale gegevensverzameling wordt opgedeeld in vier afzonderlijke delen. Elke cross-sectie kan op analoge wijze geanalyseerd worden, men besluit daarom de schattingsproblematiek allereerst alleen voor het eerste tijdstip te doorlopen (de tijdstipindex kan voorlopig achterwege blijven). Team A beheerst de theorie van hoofdstuk 3 goed en komt op grond van de klassieke testtheorie tot de volgende globale conclusies. In de eerste plaats constateert men dat de gekwadrateerde correlatie tussen de geobserveerde scores en de ware scores in de populatie, de betrouwbaarheid, wordt gegeven door

$$\rho^2_{Y\eta} = \frac{\sigma^2_\eta}{\sigma^2_Y} = \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2_\varepsilon} = \frac{25}{25 + 6.25} = .8. \qquad (10.1)$$

In de tweede plaats geldt dat de regressie van de geobserveerde toetsscore op de ware score,

$$\mathscr{E}(Y \mid \eta) = \eta, \tag{10.2}$$

lineair is. Men haalt opgelucht adem, uit (10.2) kan men concluderen dat Y, de geobserveerde score, een zuivere schatter voor η is. Hoe goed die schatter is, wordt gegeven door de betrouwbaarheid (10.1) en de schattingsfoutvariantie zal gelijk zijn aan de meetfoutvariantie σ_ε^2. Team A geeft in eerste instantie hoog op van de kwaliteiten van Y als schatter van η; deze schatter zullen in het vervolg aangeven met $\hat{\eta}$. Na enige overpeinzingen is men toch niet helemaal tevreden met deze schatter. Wat heeft men eigenlijk aan de conditionele verwachting, $\mathscr{E}(Y \mid \eta)$, als Y bekend en η onbekend is? Eigenlijk zou men de conditionele verwachting van η gegeven Y willen hebben. Verder geldt dat voor de schatting van de ware score van een individuele leerling op een meetmoment men niet over replicaties beschikt, slechts één waarneming is beschikbaar. Dit impliceert dat de zuiverheid van de geobserveerde score als schatter, op het individuele niveau bezien, niet bar veel betekent. Bij de bepaling van de verwachting, $\mathscr{E}(Y \mid \eta)$, introduceren we als gevolg van de kleine steekproef (één waarneming voor een leerling), een onzuiverheid die gelijk is aan de meetfout ε voor die ene waarneming. Ook denkt men dat er schatters te vinden zijn die een kleinere schattingsfoutvariantie hebben daar men meer informatie kan gebruiken. De verwaarloosde informatie betreft de a priori kennis met betrekking tot η, η is immers getrokken uit een bekende verdeling.

Men besluit verder te zoeken. Het punt van de verwaarloosde informatie levert gelijk een andere schatter van η op: het gemiddelde μ_η van de (marginale) verdeling van η. De schattingsfoutvariantie van deze schatter, $\tilde{\eta}$, is dan gegeven door de variantie van de (marginale) verdeling, σ_η^2. Meer algemeen kan de a priori informatie geschreven worden als

$$\eta = \mu_\eta + \zeta, \qquad \text{(a priori informatie)}$$

waarbij is ζ een meetfoutvariabele is met verwachting 0 en variantie σ_η^2.

Al snel concludeert men dat dit geen groot succes is: onzuiverheid en schattingsfoutvariantie zijn voor de a priori schatter groter dan voor de geobserveerde score schatter. Nader onderzoek leert dat deze twee schatters onafhankelijk zijn en bovendien allebei zuiver zijn in de populatie, dat wil zeggen

$$\mathscr{E}_Y(\hat{\eta}) = \mathscr{E}_Y(\tilde{\eta}) = \mu_\eta.$$

Het ligt nu voor de hand deze schatters te combineren. De optimale combinatie van twee zuivere schatters, zeg $\hat{\eta}_1$ en $\hat{\eta}_2$ met bijbehorende schattingsfoutvarianties P_1 en P_2 wordt gegeven door

$$\eta^* = P(P_1^{-1}\,\hat{\eta}_1 + P_2^{-1}\,\hat{\eta}_2), \tag{10.3}$$

waarbij P, de schattingsfoutvariantie van deze schatter, gegeven wordt door

$$P = (P_1^{-1} + P_2^{-1})^{-1}. \tag{10.4}$$

Substitutie van de a priori schatter (μ_η) en de geobserveerde score schatter (Y) en bijbehorende schattingsfoutvarianties respectievelijk σ_ε^2 en σ_η^2 in (10.3) en (10.4) levert dan

$$\eta^* = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}\,\mu_\eta + \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}\,y, \tag{10.5}$$

en

$$P = \frac{\sigma_\varepsilon^2\,\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}. \tag{10.6}$$

Deze resultaten in ogenschouw nemend herkent men hierin de Kelley-schatter voor de ware score (de kleinste-kwadratenschatter $\mathscr{E}(\eta \mid Y)$), waarmee men al bekend was uit de klassieke testtheorie (zie hoofdstuk 3). Kelley vond dit al een interessante schatter voor de ware score, daar deze schatter de gewogen som is van twee afzonderlijke schatters, één gebaseerd op de geobserveerde score van de persoon en de ander op het gemiddelde van de groep waartoe deze persoon behoort. Als de betrouwbaarheid van de toets hoog is, wordt deze schatter voornamelijk bepaald door de toetsscore Y, bij een lage betrouwbaarheid voornamelijk door het groepsgemiddelde μ_η (Lord & Novick, 1968, p. 65).
Team A is tevreden. Voor de duidelijkheid zet men de drie schatters met bijbehorende varianties van de schattingsfout nog eens onder elkaar:

$\tilde{\eta}_t = \mathscr{E}(\eta_t) = \mu_{\eta_t}$ $\qquad$ $P_t = \sigma_{\eta_t}^2$ $\qquad$ a priori schatter

$\hat{\eta}_t = \mathscr{E}(Y_t \mid \eta_t) = y_t$ $\qquad$ $P_t = \sigma_\varepsilon^2$ geobserveerde-score-schatter

$$\eta_t^* = \mathscr{E}(\eta_t \mid Y_t) = \frac{\sigma_\varepsilon^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2}\,\mu_{\eta_t} + \frac{\sigma_{\eta_t}^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2}\,y_t \qquad P_t = \frac{\sigma_\varepsilon^2\,\sigma_{\eta_t}^2}{\sigma_{\eta_t}^2 + \sigma_\varepsilon^2} \text{ Kelley-schatter}$$

Om de berekening van de schattingen van de ware scores voor de 1000 leerlingen in de steekproef op de vier tijdstippen te vereenvoudigen, maakt men gebruik van tabel 10.1.

Tabel 10.1
Schatters en schattingsfoutvarianties voor de vier tijdstippen

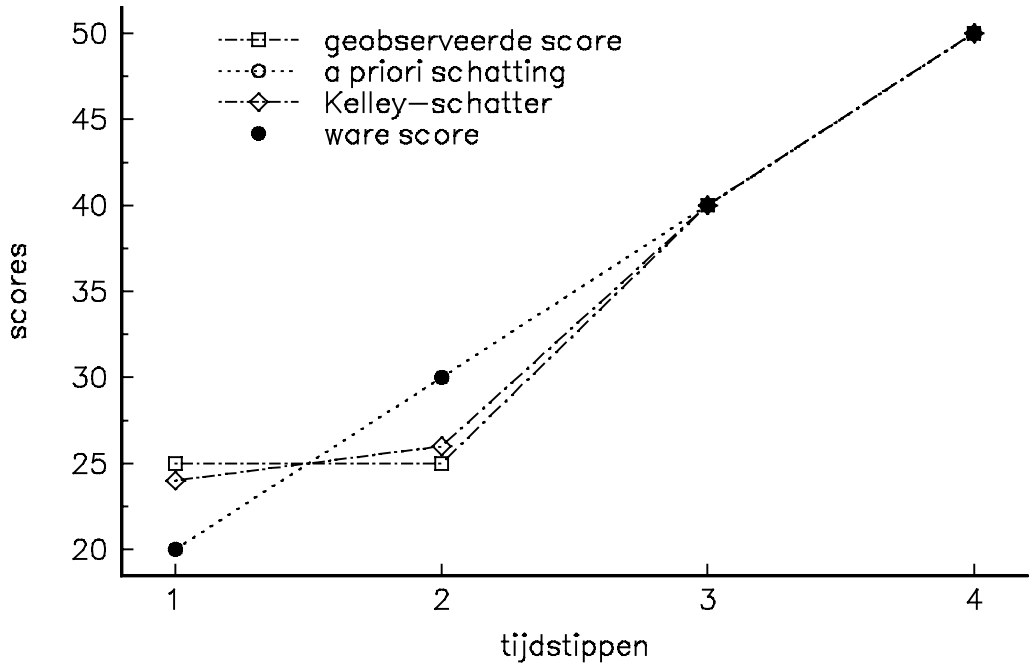
	a priori		geobserveerde score		Kelley	
tijdstip	$\tilde{\eta}$	P	$\hat{\eta}$	P	η^*	P
1	20	25	y_1	6.25	$4+.8y_1$	5
2	30	25	y_2	6.25	$6+.8y_2$	5
3	40	25	y_3	6.25	$8+.8y_3$	5
4	50	25	y_4	6.25	$10+.8y_4$	5

Om enig inzicht te verkrijgen in het functioneren van deze drie schatters, besluit men om voor twee leerlingen het gedrag van deze schatters te onderzoeken. Er van uitgaande dat leerling A op alle vier de tijdstippen een ware score heeft die gelijk is aan het populatiegemiddelde (ware scores: 20, 30, 40 en 50), creëert men de volgende observaties voor de vier tijdstippen: 25, 25, 40 en 50. De toegevoegde meetfout is dus respectievelijk: 5, -5, 0 en 0. In figuur 10.2 zijn de ware scores en de drie besproken schattingen van de ware scores voor leerling A weergegeven voor de vier tijdstippen.

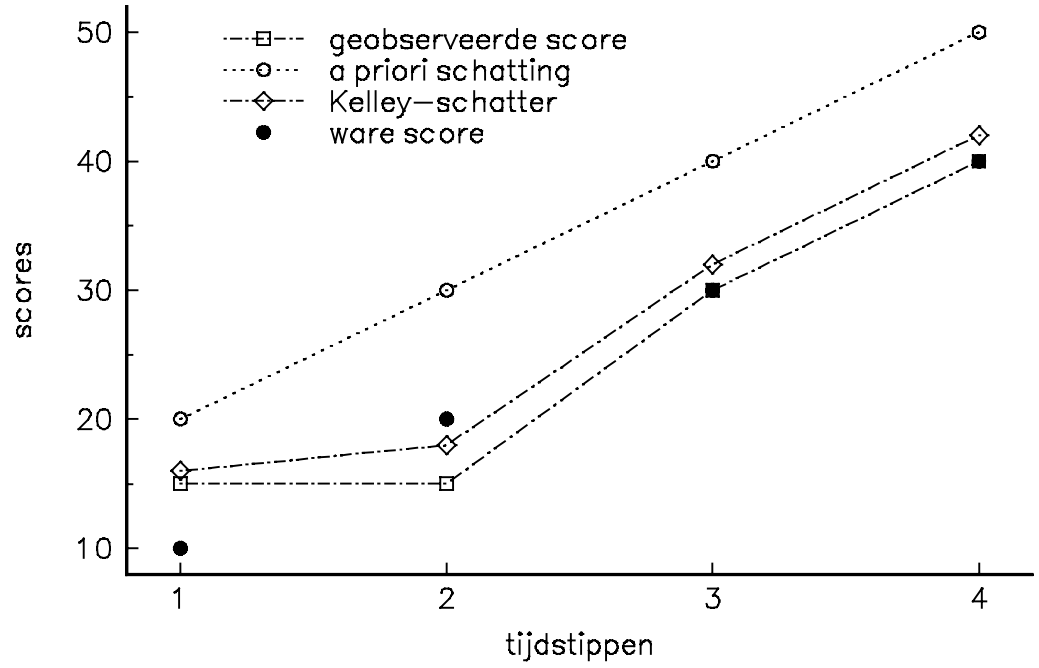
In de eerste plaats kunnen we in figuur 10.2 constateren dat de a priori schatting op alle tijdstippen samenvalt met de ware score, niet zo verwonderlijk als men zich realiseert dat de a priori schatting de gemiddelde ware score in de populatie is. Op tijdstip 3 en 4 vallen ook de geobserveerde score schattingen samen met de respectievelijke ware scores, ook niet opzienbarend daar de toegevoegde meetfout op dat tijdstip 0 was. Omdat de a priori schatting en geobserveerde-score-schatting voor tijdstip 3 en 4 samenvallen, resulteren ook de Kelley-schattingen in de ware scores voor leerling A. De geobserveerde-score-schattingen op tijdstip 1 en 2 zitten er behoorlijk naast, de mate waarin is bepaald door de toegevoegde meetfout, dat is respectievelijk plus en minus 2 × de standaardafwijking van de meetfout. Op tijdstip 1 en 2 functioneert de Kelley-schatter beter dan de geobserveerde-score-schatter, de Kelley-schatter duwt (Engels: 'shrinkage') de geobserveerde scores in de richting van de a priori schatter en

komt zodoende dichter in de buurt van de ware scores. Hoe hard de Kelley-schatter duwt, wordt bepaald door de betrouwbaarheid van de observaties (zie tabel 10.1).

De ware scores voor leerling B zijn respectievelijk 10, 20, 30 en 40. De toegevoegde meetfout is respectievelijk: 5, -5, 0 en 0, hetgeen resulteert in de geobserveerde scores 15, 15, 30 en 40. In figuur 10.3 zijn de ware score schattingen weergegeven voor leerling B.



Figuur 10.2
Schattingen van de ware scores voor de 'gemiddelde' leerling A



Figuur 10.3 Schattingen van de ware scores voor leerling B

De a priori schattingen zitten er behoorlijk naast, en wel 10 scorepunten. Het verschil tussen de geobserveerde scores en de ware scores bij leerling B is hetzelfde als bij leerling

A en is gelijk aan de toegevoegde meetfout op de 4 momenten, respectievelijk 5, 5,0 en 0. Ook hier duwt de Kelley-schatter de geobserveerde scores in de richting van de a priori schat- ter. Op tijdstip 1, 3 en 4 is het effect hiervan dat de afstand tussen de ware score en de Kelley-schatting groter is dan die tussen de ware score en de geobserveerde score. Op tijdstip 2 geldt het omgekeerde.

Uit deze twee voorbeelden kunnen we concluderen dat geen van de drie besproken schatters het onder alle omstandigheden goed doet. Afhankelijk van de relatieve positie van een leerling in de populatie en de grootte van de meetfout, gaat de voorkeur uit naar een van de drie schatters. Welke schatter over individuen heen het predikaat 'best' verdient, zullen we bespreken nadat de dynamische benadering besproken is.

10.2.3 Dynamische benadering

Ook team B begint met een inspectie van de meetvergelijking in de klassieke testtheorie, maar beperkt zich in eerste instantie tot één meetmoment. Men realiseert

zich dat de meetvergelijking in de klassieke testtheorie de relatie beschrijft tussen toevalsvariabelen in de populatie. Met deze constatering als uitgangspunt gaat men het schattingsprobleem van de ware score voor een bepaald individu specificeren. De meetvergelijking in de klassieke testtheorie beschrijft niets anders dan de relatie tussen de toevalsvariabelen Y en η in een populatie, met een niet gespecificeerde gezamenlijke verdeling. De observeerbare variabele Y is in dit geval behept met een meetfout. Intuïtief is het duidelijk dat de meting van Y ons iets kan leren over η. Of, anders gezegd, stel dat we over a priori informatie over η beschikten, dan zou kennis van Y deze informatie omtrent η moeten verbeteren. De volgende vraag is nu relevant: "Gegeven de observatie $Y = y$, wat is dan de beste schatting van η?" Eerst geven we inhoud aan het concept 'best'. Een veel gebruikt criterium hiervoor is dat van de kleinste-kwadraten. Hierbij wordt gezocht naar een schatter $\eta^*(Y)$ die een functie is van de meting $Y = y$ zodanig dat

$$\mathcal{E}[\eta - \eta^*(Y)]^2 \leq \mathcal{E}[\eta - g(Y)]^2, \tag{10.7}$$

voor elke functie g. De oplossing van (10.7) wordt gegeven door

$$\eta^*(Y) = \mathcal{E}(\eta \mid Y).$$

Merk nu op dat $\eta^*(Y)$ een toevalsvariabele is, in tegenstelling tot de realisatie $\eta^*(y)$ daar- van voor observatie $Y = y$. Problematisch is dat $\eta^*(Y)$ meestal niet een lineaire functie van Y is. Daarnaast beschikken we in de klassieke testtheorie meestal niet over de gezamenlijke verdeling van η en Y, zodat het onmogelijk is om $\mathcal{E}(\eta \mid Y)$ te bepalen. Daarom zullen we een extra aanname maken. We veronderstellen namelijk dat $\eta^*(Y)$ een lineaire functie van Y is,

$$\eta^*(Y) = aY + b \tag{10.8}$$

waarbij a en b te bepalen constanten zijn. De oplossing van (10.8), onder de restrictie van vergelijking (10.7), is gegeven door:

$$a = \frac{\sigma_{Y\eta}}{\sigma_\eta^2} \tag{10.9}$$

en

$$b = \mu_\eta - \frac{\sigma_{Y\eta}}{\sigma_Y^2}\,\mu_Y, \tag{10.10}$$

waarbij $\sigma_{Y\eta}$ de covariantie tussen Y en η is. Substitutie van (10.9) en (10.10) in (10.8) levert dan de beste lineaire schatter van η gebaseerd op Y:

$$\eta^*(Y) = \mu_\eta - \frac{\sigma_{Y\eta}}{\sigma_Y^2}\mu_Y + \frac{\sigma_{Y\eta}}{\sigma_Y^2} Y. \tag{10.11}$$

De variantie van de schattingsfout is gegeven door

$$P = \mathscr{E}[\eta - \eta^*(Y)]^2 = \sigma_\eta^2 - \frac{\sigma_{Y\eta}^2}{\sigma_Y^2}. \tag{10.12}$$

Het geoefende oog van team B herkent in (10.11) en (10.12) natuurlijk de Kelley-schatter met bijbehorende schattingsfoutvariantie (herschrijf (10.5) en (10.6) en maak hierbij gebruik van de formules uit de klassieke testtheorie). Daar in dit voorbeeld de ware vaardigheidsverdeling multivariaat normaal en de meetfout normaal verdeeld is, is ook de conditionele verdeling van η gegeven Y normaal verdeeld, waarbij het gemiddelde gegeven wordt door (10.11) en de variantie door (10.12).
Nu men het schattingsprobleem in essentie voor twee toevalsvariabelen heeft opgelost gaat men dit toepassen in een longitudinale context. De subscripten bij de variabelen die in het vervolg gebruikt worden geven nu de tijdstippen weer. Op het eerste meetmoment lijkt de Kelley-schatter en schattingsfoutvariantie de aangewezen keus, dus

$$\begin{aligned} \eta_1^* &= \mu_{\eta_1} - \frac{\sigma_{Y_1\eta_1}}{\sigma_{Y_1}^2}\mu_{Y_1} + \frac{\sigma_{Y_1\eta_1}}{\sigma_{Y_1}^2} Y_1, \\ P_1 &= \sigma_{\eta_1}^2 - \frac{\sigma_{Y_1\eta_1}^2}{\sigma_{Y_1}^2}. \end{aligned} \tag{10.13}$$

In tegenstelling tot team A onderkent team B dat, gegeven de knikkervaardigheidsverdeling in de populatie, het mogelijk is η_2 te voorspellen met η_1. Inmiddels weten we hoe dat moet en de oplossing wordt gegeven door

$$\eta_{2|1}^* = \mathscr{E}(\eta_2 \mid \eta_1) = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma_{\eta_1}^2}\mu_{\eta_1} + \frac{\sigma_{\eta_1\eta_2}}{\sigma_{\eta_1}^2}\eta_1. \tag{10.14}$$

In de praktijk beschikken we niet over η_1; we zullen ons tevreden moeten stellen met een schatting hiervan, zeg η_1^*. Voorspellen is nu niets anders dan substitutie van deze schatting (10.13) in (10.14) hetgeen resulteert in:

$$\eta^*_{2|1} = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\mu_{\eta_1} + \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\eta^*_1,$$

ofwel

$$\eta^*_{2|1} = \mu_{\eta_2} - \frac{\sigma_{\eta_1\eta_2}}{\sigma^2_{\eta_1}}\,\frac{\sigma_{y_1\eta_1}}{\sigma^2_{y_1}}\,(y_1 - \mu_{y_1}). \tag{10.15}$$

De berekening van de variantie van (10.15) gaat recht toe recht aan en levert op:

$$P_{2|1} = \sigma^2_{\eta_2} - \frac{\sigma^2_{\eta_1\eta_2}\sigma^2_{y_1\eta_1}}{\sigma^4_{\eta_1}\sigma^2_{y_1}}. \tag{10.16}$$

Deze voorspelling en schattingsfoutvariantie zijn in wezen niets anders dan de a priori informatie met betrekking tot η_2 gegeven de waarneming y_1 op tijdstip 1. Merk op dat deze a priori informatie in feite een a priori verdeling voor η_2 is met gemiddelde $\eta^*_{2|1}$ en variantie $P_{2|1}$, die in ons voorbeeld normaal verdeeld is. Als we op tijdstip 2 deze a priori informatie in het dynamische geval vergelijken met de a priori informatie bij de statische benadering, dan valt op dat het gemiddelde μ_{η_2} in het dynamische geval gecorrigeerd wordt (vergelijking 10.15) en dat de variantie $\sigma^2_{\eta_2}$ verkleind wordt (zie 10.16). Met andere woorden, onze a priori informatie op tijdstip 2 wordt meer specifiek voor een individu, daar we immers rekening houden met de geobserveerde score y_1 van dit individu. Bovendien is de hoeveelheid informatie groter, zodat de onzekerheid over iemands positie in de populatie afneemt.

Als we het meetresultaat op tijdstip 2, y_2, willen combineren met de a priori kennis op tijdstip 2, dan kan dat beschreven worden als het combineren van twee schatters (zie ook paragraaf 10.2.2 voor de combinatie van een a priori schatter en de geobserveerde-score-schatter) of, analoog aan hierboven, door het bepalen van de conditionele verwachting $\mathscr{E}(\eta_2 \mid Y_1, Y_2)$. Beide resulteren in de volgende schatting voor η_2:

$$\eta^*_2 = \eta^*_{2|1} + K_2(y_2 - \eta^*_{2|1}),$$

waarbij K_2 gegeven is door:

$$K_2 = P_{2|1}(P_{2|1} + \sigma^2_\varepsilon)^{-1}.$$

De bijbehorende schattingsfoutvariantie, P_2, wordt gegeven door:

$$P_2 = P_{2|1} - K_2 P_{2|1}.$$

De bepaling van een schatter voor η_3 gaat analoog aan de procedure voor η_2. Voorspel η_3 met behulp van η_2, vul de lopende schatting van η_2 in deze vergelijking in en combineer deze predictie met de observatie y_3 op het derde tijdstip. Uiteraard kunnen we zo doorgaan voor de volgende tijdstippen. Merk op dat we voor de voorspelling van η_3 alleen η_2 gebruiken en niet η_1. Met andere woorden, we gaan ervan uit dat, gegeven η_2, η_1 ons niets meer kan leren over η_3. Of anders gezegd, de partiële correlatie tussen η_1 en η_3 veronder- stellen we gelijk aan nul. Dat geldt ook op de andere tijdstippen, dus alle partiële correlaties tussen de latente variabelen zijn 0, behalve voor aanliggende tijdstippen. Dit impliceert dat de covariantiematrix van η een bepaalde structuur heeft, die in de literatuur aangeduid wordt met 'autoregressief van de eerste orde'. De hier beschreven recursieve schattingsprocedure staat bekend als het Kalmanfilter, de schattingen als Kalmanfilterschattingen.

Team B is tevreden met het resultaat. Men signaleert echter één minpunt. Men realiseert zich dat de Kalmanfilterschattingen voor de vier tijdstippen niets anders zijn dan de conditionele verwachtingen: $\mathcal{E}(\eta_1 \mid y_1)$, $\mathcal{E}(\eta_2 \mid y_1, y_2)$, $\mathcal{E}(\eta_3 \mid y_1, y_2, y_3)$ en $\mathcal{E}(\eta_4 \mid y_1, y_2, y_3, y_4)$. Bezien we deze reeks, dan kan geconstateerd worden dat het aantal waarnemingen waarop deze conditionele verwachtingen gebaseerd zijn in de tijd toeneemt. Op het eerste tijdstip gebruiken we slechts één waarneming, terwijl op het vierde tijdstip gebruik gemaakt is van alle meetresultaten. Beschikken we over vier waarnemingen, dan geldt alleen voor de Kalmanfilterschatting op het vierde tijdstip dat alle informatie uit de data verwerkt is in de schatter. Voor de Kalmanfilterschatting op tijdstip 3, bijvoorbeeld, hebben we geen gebruik gemaakt van de laatste waarneming. Het ligt dus voor de hand die informatie alsnog toe te voegen, dat is, door $\mathcal{E}(\eta_3 \mid y_1, y_2, y_3, y_4)$ te bepalen. Voor de Kalmanfilterschattingen op tijd- stip 2 en 1, berekenen we dan respectievelijk $\mathcal{E}(\eta_2 \mid y_1, y_2, y_3, y_4)$ en $\mathcal{E}(\eta_1 \mid y_1, y_2, y_3, y_4)$. De conditionele verwachting van η, op een tijdstip gegeven alle data duidt men aan met de naam gladgestreken Kalmanfilterschatting. Het bepalen van de gladgestreken schattingen kan eenvoudig geïllustreerd worden aan het kleinste-kwadratenprobleem in het begin van deze paragraaf. Daar zochten we de conditionele verwachting van η gegeven Y. Maar dit is in wezen niets anders dan de univariate regressie van η op Y. Stel dat we de multivariate lineaire regressie bepalen van de vector η op de vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)'$, dat is

$$\hat{\eta} = \mathcal{E}(\eta \mid \boldsymbol{Y}) = \boldsymbol{\mu}_\eta + \Sigma_{\boldsymbol{Y}\eta}\, \Sigma_{\boldsymbol{Y}}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}}), \tag{10.17}$$

dan beschikken we in een keer over de gladgestreken schattingen in de vector $\hat{\eta}$. De covariantiematrix van de gladgestreken schattingen is

$$\boldsymbol{P} = \Sigma_{\eta} - \Sigma_{\boldsymbol{Y}\eta}\,\Sigma_{\boldsymbol{Y}}^{-1}\,\Sigma_{\eta}. \qquad (10.18)$$

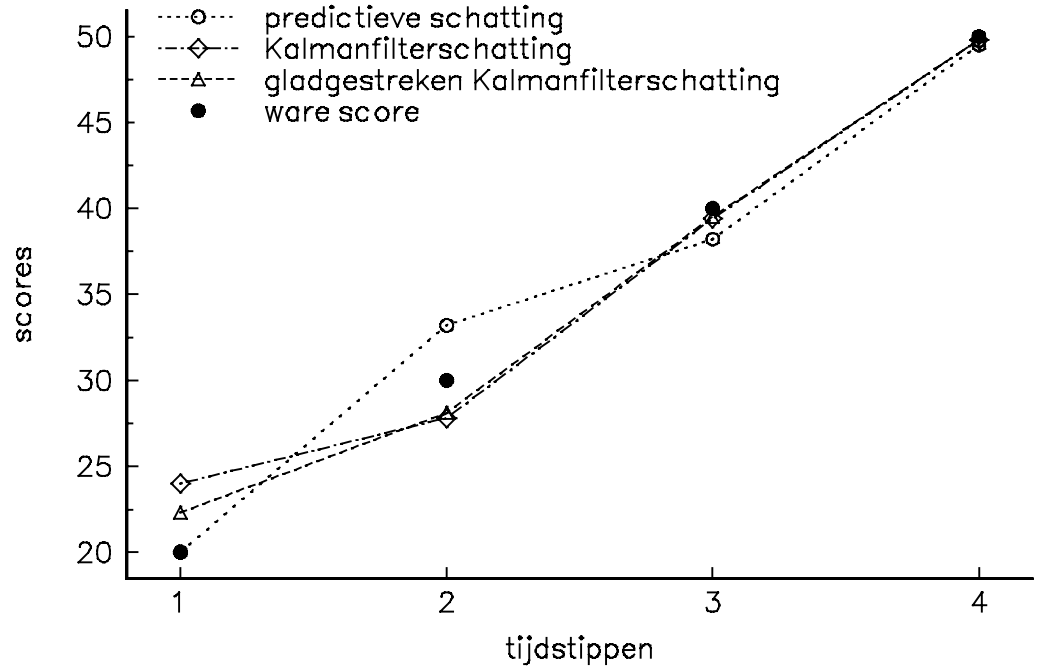
Merk op dat voor de klassieke testtheorie geldt dat de covariantiematrix tussen de vectoren $\boldsymbol{Y}$ en η, $\Sigma_{\boldsymbol{Y}\eta}$, gelijk is aan de variantie-covariantiematrix van de vector η, dat wil zeggen $\Sigma_{\boldsymbol{Y}\eta} = \Sigma_{\eta}$.

Tabel 10.2
Schatters en schattingsfoutvarianties voor de vier tijdstippen

	predictie		Kalmanfilter		gladgestreken Kalmanfilter	
tijd-stip	$\eta^*_{t\mid t-1}$	$P_{t\mid t-1}$	η^*_t	P_t	$\hat{\eta}_t$	P_t
1	20	25	$4 + .8y_1$	5	$\eta^*_1 + .33(\hat{\eta}_2 - 14 - .8\eta^*_1)$	4.06
2	$14 + .8\eta^*_1$	12.20	$\eta^*_{2\mid 1} + .66(y_2 - \eta^*_{2\mid 1})$	4.13	$\eta^*_2 + .28(\hat{\eta}_3 - 16 - .8\eta^*_2)$	3.47
3	$16 + .8\eta^*_2$	11.65	$\eta^*_{3\mid 2} + .65(y_3 - \eta^*_{3\mid 2})$	4.07	$\eta^*_3 + .28(\hat{\eta}_4 - 18 - .8\eta^*_3)$	3.47
4	$18 + .8\eta^*_3$	11.60	$\eta^*_{4\mid 3} + .65(y_4 - \eta^*_{4\mid 3})$	4.06	η^*_4	4.06

Een recursieve procedure (nu achterwaarts) voor het berekenen van de gladgestreken schat- tingen, waarin alleen gebruik gemaakt wordt van de predictieve filterschattingen en Kalman- filterschattingen met bijbehorende covarianties, staat vermeld in Jazwinski (1970).

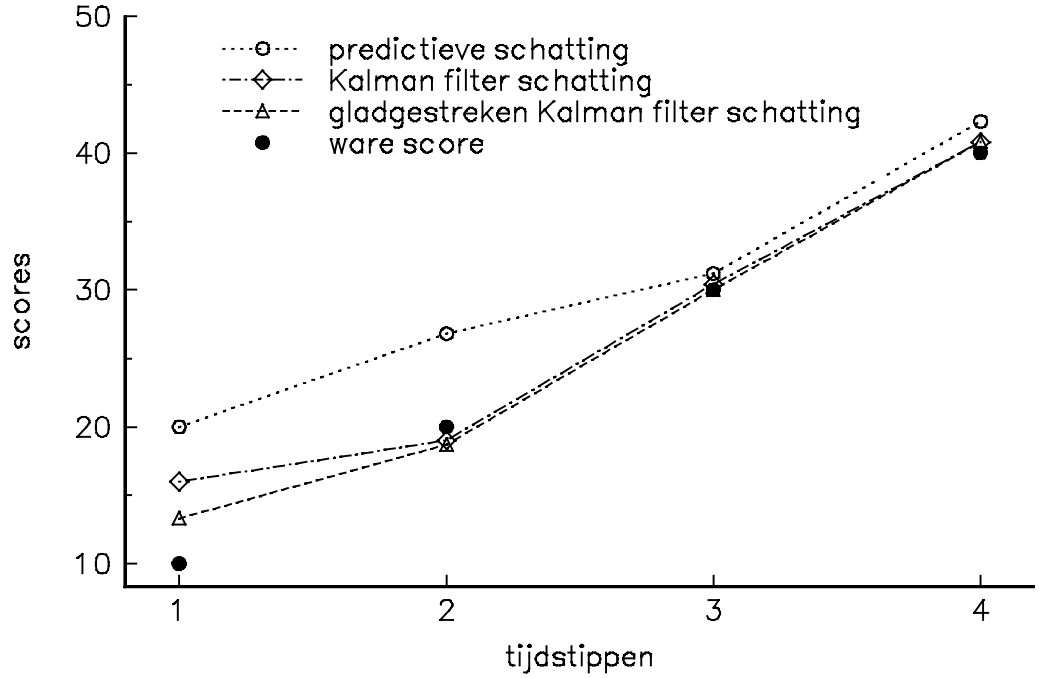
Ook team B gaat de ware scores uitrekenen voor de 1000 leerlingen in de steekproef. In tabel 10.2 zijn zijn de resultaten voor de predictie-, de Kalmanfilter- en de gladgestreken Kalmanfilterschattingen op de vier tijdstippen vermeld. Tenslotte kijkt team B naar het functioneren van de door hen geconstrueerde schatters.



Figuur 10.4 Schattingen van de ware scores voor de 'gemiddelde' leerling A

Om het gedrag van de schatters te onderzoeken maakt men, net als bij team A, gebruik van de scores van leerling A en leerling B (zie paragraaf 10.2.2). In figuur 10.4 zijn de resultaten voor leerling A weergegeven. Op het eerste tijdstip is de a priori kennis bij de statische en dynamische aanpak even groot, met uitzondering van de gladgestreken Kalmanfilterschatting. De reden hiervoor is dat de a priori schatting en de predictieve schatting samenvallen, dus ook de Kelley-schatting en de Kalmanfilterschatting. Merk op dat de gladgestreken schatting op het eerste meetmoment en in mindere mate op het tweede tijdstip het dichtst komt bij de ware score. Voor deze leerling kan de informatie uit de latere tijdstippen dus de schattingen op de eerste twee tijdstippen tot op zekere hoogte in de goede richting corrigeren. Kijken we naar de schattingen voor leerling B (zie figuur 10.5), dan valt op dat de predictieve schattingen op de laatste drie meetmomenten dichter bij de ware scores liggen dan de a priori schattingen in het statische geval.

Dit heeft tot gevolg dat de Kalmanfilterschattingen op deze momenten de ware scores beter benaderen dan de Kelley-schattingen bij de statische benadering. Het plaatje is wederom het fraaist voor de gladgestreken schattingen. Deze schattingen, komen over de vier tijdstippen bezien, immers het dichtst bij de ware scores.



Figuur 10.5

Schattingen van de ware scores voor leerling B

10.2.4 Evaluatie statische en dynamische benadering

Het wordt tijd om de door team A en team B voorgestelde schatters te evalueren. Beide teams hebben voor de 1000 leerlingen in de steekproef op alle vier de tijdstippen schattingen en bijbehorende schattingsfoutvarianties uitgerekend en ter evaluatie aan de heer Knikker aan- geboden. Om de schatters te kunnen evalueren, zullen we eerst enige criteria moeten aan- nemen waarop de evaluatie van de schatters kan plaatsvinden. De heer Knikker besteedt deze klus uit aan een statisticus, aan wie hij alle materiaal, inclusief de ware scores, beschikbaar stelt. Deze statisticus ziet twee mogelijke manieren om de zaak te evalueren. In de eerste plaats kan hij de schatters beoordelen op hun statistische eigenschappen. Omdat alle gegevens beschikbaar zijn, kan hij ook de schattingen en de ware scores van alle 1000 leerlingen vergelijken; dit is de tweede manier.

We bekijken eerst de statistische eigenschappen. In de eerste plaats valt op dat alle voorgestelde schatters, zowel die van team A als die van team B, kleinste-kwadraten-schatters zijn, die alleen verschillen in de mate waarin ze de beschikbare informatie gebruiken. De volgende tabel 10.3 vat de bron en de hoeveelheid informatie voor de diverse schatters samen. De bron van de informatie refereert aan het meetmodel en het groeimodel, terwijl de hoeveelheid informatie het aantal tijdstippen aanduidt.

Tabel 10.3

Hoeveelheid informatie van de diverse schatters uitgesplitst naar bron

	bron informatie	
	groeimodel	meetmodel
schatter op tijdstip t	η_t	y_t
a priori	t	geen
geobserveerde score	geen	t
Kelley	t	t
predictieve	1 t/m t	1 t/m $t-1$
Kalmanfilter	1 t/m t	1 t/m t
gladgestreken Kalmanfilter	alle	alle

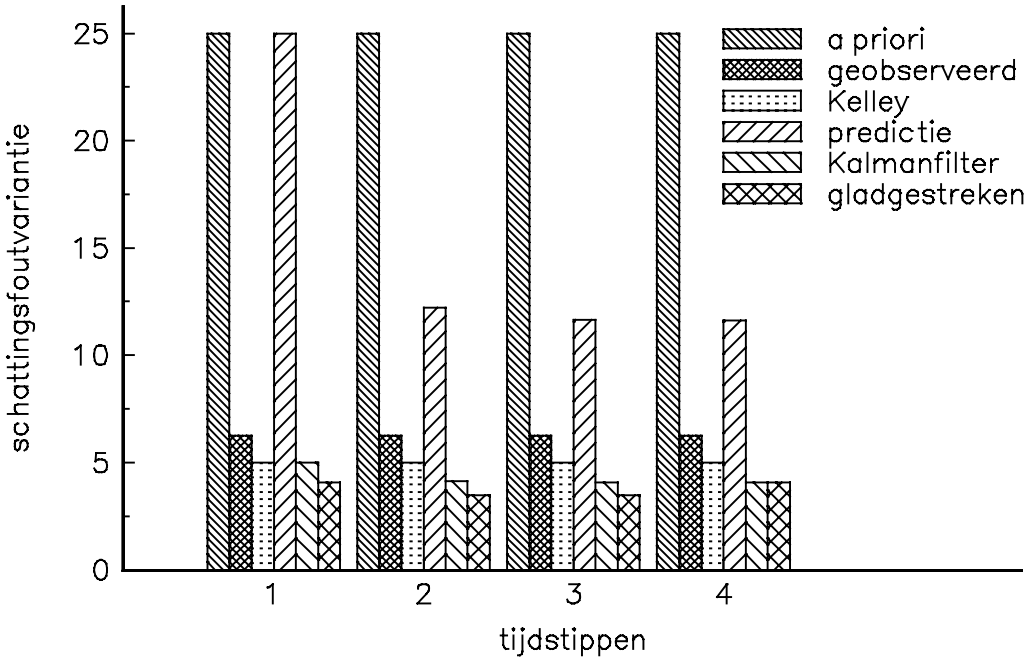
Naarmate een schatter meer informatie gebruikt is de schattingsfoutvariantie kleiner, zoals uit de statistiek bekend is. De schattingsfoutvariantie, als indicatie voor de zekerheid van de schatting, is dan ook het eerste criterium om de schatters te vergelijken. Merk op dat, met de klassieke testtheorie als meetmodel, alle schattingsfoutvarianties op voorhand bekend zijn zonder ook maar een observatie gedaan te hebben, dat is als men de relatie kent tussen de toevalsvariabelen η_t en Y_t. In figuur 10.6 zijn met behulp van staafdiagrammen op de vier tijdstippen de schattingsfoutvarianties van de zes besproken schatters grafisch weergegeven.

We vergelijken eerst de schattingsfoutvarianties van de drie cross-sectionele schatters. De schattingsfoutvarianties van de afzonderlijke schatters zijn over de vier tijdstippen gelijk (gelijke betrouwbaarheid voor elk tijdstip). De kleinste schattingsfoutvariantie heeft de Kelley- schatter (5), gevolgd door de geobserveerde-score-schatter (6.25) en de a priori schatter (25). In het algemeen kan men zeggen dat van de cross-sectionele schatters de Kelley-schatter altijd de kleinste variantie heeft.

De Kelley-schatter gebruikt immers alle cross-sectionele informatie. De betrouwbaarheid van

de toets bepaalt de volgorde van de andere twee cross-sectionele schatters. Is de betrouwbaarheid groter dan .5, dan heeft de geobserveerde-score-schatter een kleinere variantie dan de a priori schatter; het omgekeerde geldt als de betrouwbaarheid kleiner is dan .5. Kijken we vervolgens naar de dynamische benadering, dan zien we dat de gladgestreken Kalmanfilterschatter op alle tijdstippen de kleinste schattingsfoutvariantie heeft, behoudens op het laatste tijdstip waarop deze schatter gelijk is aan de

Kalmanfilterschatter. Ook zien we dat de Kalmanfilterschatters het beter doen dan de predictieve schatters. Dit is logisch, daar de eerstgenoemde schatter in vergelijking met de predictieve schatter een extra waarneming, dat wil zeggen, extra informatie gebruikt. De orde van grootte van de schattingsfoutvariantie van de predictieve schatter hangt natuurlijk af van de mate waarin we in staat zijn de latente vaardigheid te voorspellen op een volgend tijdstip. Een maat hiervoor is de gekwadrateerde correlatie tussen de latente vaardigheden op twee tijdstippen. Een vergelijking van de schattingsfoutvarianties van de statische en dynamische schatters leert ons dat de statische equivalenten van de dynamische schatters een beduidend grotere schattingsfoutvariantie hebben. Hoe groot de verschillen zijn, hangt in het algemeen af van de toetsbetrouwbaarheid en van de mate waarin de latente vaardigheid voorspeld kan worden.

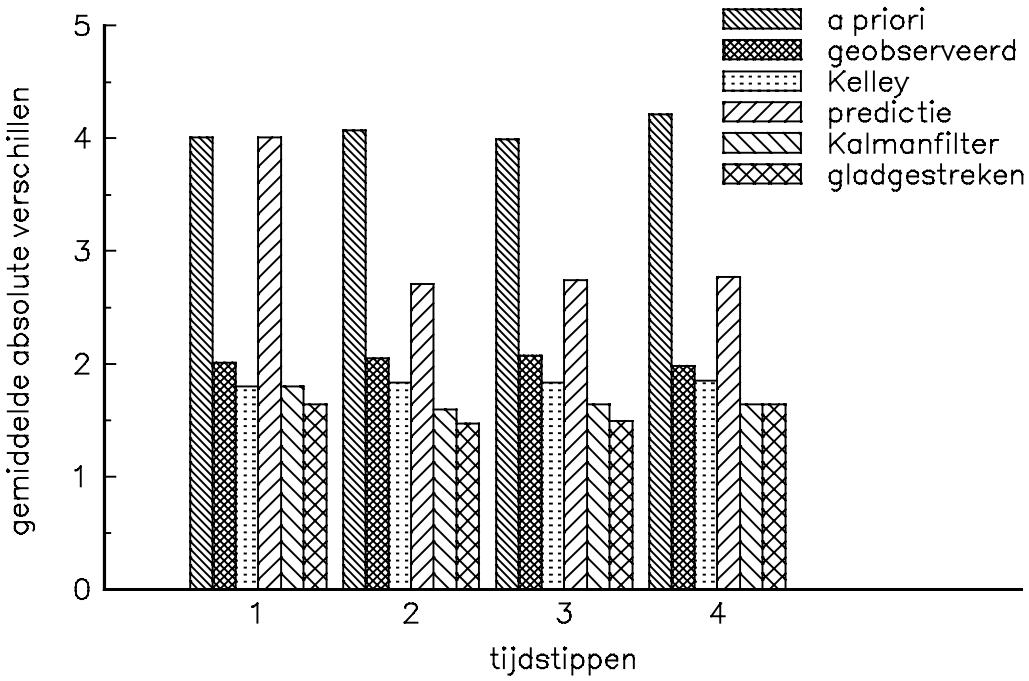


Figuur 10.6
Histogram voor de diverse schattingsfoutvarianties voor de vier tijdstippen

De tweede statistische eigenschap om schatters te beoordelen is de zuiverheid van schatters. Alle besproken schatters zijn zuiver in de populatie terwijl de geobserveerde-score-schatter bovendien zuiver is voor een individu. Aan deze laatstgenoemde vorm van zuiverheid hebben we echter niet zoveel, aangezien we op een tijdstip voor een individu meestal niet over replicaties beschikken. Wel kan deze eigenschap van de geobserveerde-score-schatter handig zijn voor het berekenen van groepsgemiddelden. Denk hierbij bijvoorbeeld aan een apart gemiddelde voor jongens en meisjes.

De statisticus concludeert dat op het criterium zuiverheid de schatters elkaar in wezen niet ontlopen en besluit daarom, het criterium zuiverheid niet te laten meewegen en zich alleen te beperken tot de schattingsfoutvariantie.

Een tweede evaluatiemogelijkheid behelst het vergelijken van de schattingen en de ware scores in de steekproef. Twee criteria om de schatters te beoordelen, acht de statisticus zinvol



Figuur 10.7

Histogram gemiddelde absolute verschil ware scores en diverse schattingen voor de vier tijdstippen

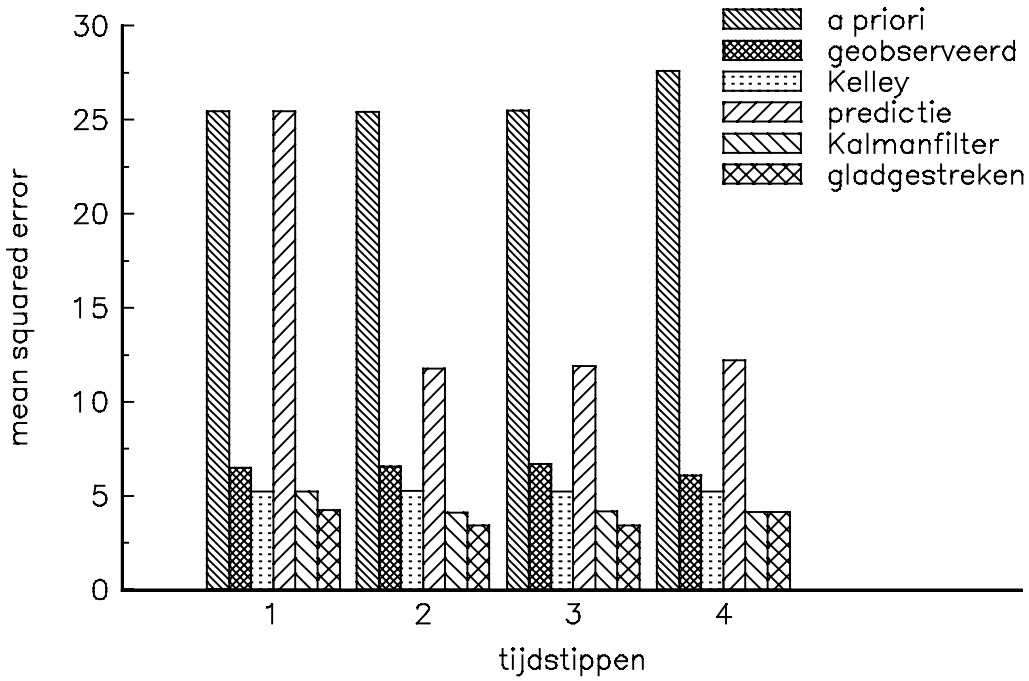
in dit verband: het gemiddelde absolute verschil en het gemiddelde gekwadrateerde verschil (’Mean Squared Errors’). In figuur 10.7 is voor elk tijdstip het gemiddelde absolute verschil tussen ware score en schatting voor de diverse schatters weergegeven, en in figuur 10.8 het gemiddelde gekwadrateerde verschil.

De conclusies aangaande de rangorde van de schatters is niet anders dan bij de bespreking van de schattingsfoutvarianties. Dit is niet zo verwonderlijk als men zich realiseert dat voor grote steekproeven de MSE gelijk zal zijn aan de schattingsfoutvariantie. Bovendien hebben absolute verschillen en gekwadrateerde verschillen een hoop gemeen.

De statisticus komt tot de volgende conclusies aangaande de analyses van de psychometrici. Als men kiest voor momentopnamen, dat is de statische benadering, dan is de Kelley-schatter aan te bevelen. Kiest men een dynamische aanpak terwijl men

bovendien over de data van alle tijdstippen beschikt, dan is de gladgestreken Kalmanfilterschatter de aangewezen keus. Wil men echter tussentijds al over schattingen kunnen beschikken, de meest voorkomende situatie, dan is de Kalmanfilterschatting te prefereren. Heeft men longitudinale data, kies dan ook voor een dynamische aanpak. De winst die een dynamische benadering oplevert, kan erg groot zijn.

Knikker vindt de resultaten redelijk. Toch constateert hij dat de psychometrici er soms behoorlijk naast zitten. Afhankelijk van de gekozen schatter zitten zij er gemiddeld gezien ongeveer 1.5 tot 4 punten naast op de knikkervaardigheidsschaal. Ook verbaast het Knikker, dat de schattingsfoutvarianties van de diverse schatters, hoewel van verschillende grootte, voor elke leerling gelijk zijn. Knikker verwachtte namelijk dat het vaardigheidsniveau van sommige leerlingen nauwkeuriger geschat zou kunnen worden dan dat van andere leerlingen.



Figuur 10.8
Histogram 'Mean Squared Errors' (MSE) voor de vier tijdstippen

Tenslotte vraagt Knikker zich af of de resultaten anders geweest zou zijn als hij niet alle informatie ter beschikking had gesteld. Hij had de psychometrici bijvoorbeeld alleen de geobserveerde toetsscores kunnen verschaffen en niet de informatie over de populatie. Aangaande dit laatste punt kunnen de psychometrici Knikker gerust stellen. Onder zekere assumpties en restricties is het mogelijk de gegevens van de populatie te

achterhalen. Een methode om de populatieparameters te schatten staat beschreven in de volgende paragraaf.

10.2.5 Schattingen van structurele parameters

In het voorbeeld van de knikkervaardigheid was het uitgangspunt dat alle parameters behalve de ware scores bekend waren. In de praktijk zal dat niet zo zijn en zullen de parameters uit de observaties geschat moeten worden. Dit is mogelijk door de individuele tijdreeksen te beschouwen als replicaties van een onderliggende tijdreeks op populatieniveau. Hoe het een en ander zijn beslag krijgt, kan het beste geïllustreerd worden aan de hand van het zogenaamde simplexmodel. Het simplexmodel is een model met een bepaalde covariantiestructuur die vaak van toepassing is op longitudinale data. Hierbij wordt dezelfde variabele bij dezelfde individuen op verschillende tijdstippen gemeten, of in een situatie waarbij de variabelen niet geordend zijn in de tijd maar bijvoorbeeld naar toenemende complexiteit. Een voorbeeld van laatstgenoemde situatie kan men vinden bij Guttman (1954) voor spreekvaardigheid. De typische structuur van simplexmodellen, in de correlatiematrix nemen de correlaties van de diagonaal af gezien af, wordt gegenereerd door een onderliggend eerste-orde-autoregressief proces. Voor een uitvoerige introductie van deze modellen verwijzen we naar Guttman (1954), Jöreskog (1970) en Imbos (1989).

De schattings- en identificatieproblematiek van de parameters van het simplexmodel bespreken we in het kort. Omwille van de eenvoud beperken we ons tot gestandaardiseerde metingen op vier tijdstippen, y_t $(t = 1, 2, 3, 4)$. Het meetmodel op de vier tijdstippen kan wederom beschreven worden met de meetvergelijking uit de klassieke testtheorie

$$y_t = \eta_t + \varepsilon_t \qquad t = 1, 2, 3, 4.$$

Het groeimodel heeft een autoregressieve structuur die met de volgende drie vergelijkingen beschreven kan worden

$$\eta_t = \beta_t \eta_{t-1} + \zeta_t \qquad t = 2, 3, 4. \tag{10.19}$$

In (10.19) kan β_t geïnterpreteerd worden als de regressiecoëfficiënt van η_t op η_{t-1} en ζ_t als de meetfout met bijbehorende variantie Ψ_t (het onverklaarde deel van de variantie van η_t). Merk op dat de latente variabelen η_t en de geobserveerde variabelen y_t op dezelfde schaal liggen, zodat bij gestandaardiseerde metingen geldt dat, voor alle t,

$\mathcal{E}(\eta_t) = \mathcal{E}(y_t) = 0$. De correlatiematrix Σ_y van de geobserveerde variabelen heeft de volgende vorm:

$$\sum_y = \begin{pmatrix} \sigma^2_{\eta_1} + \sigma^2_{\varepsilon_1} & & & \\ \beta_2 \sigma^2_{\eta_1} & \sigma^2_{\eta_2} + \sigma^2_{\varepsilon_2} & & \\ \beta_2 \beta_3 \sigma^2_{\eta_1} & \beta_3 \sigma^2_{\eta_2} & \sigma^2_{\eta_3} + \sigma^2_{\varepsilon_3} & \\ \beta_2 \beta_3 \beta_4 \sigma^2_{\eta_1} & \beta_3 \beta_4 \sigma^2_{\eta_2} & \beta_4 \sigma^2_{\eta_3} & \sigma^2_{\eta_4} + \sigma^2_{\varepsilon_4} \end{pmatrix},$$

waarbij $\sigma^2_{\eta_t} = \beta^2_t \sigma^2_{\eta_{t-1}} + \Psi_t \quad (t = 2, 3, 4)$. Het blijkt dat niet alle parameters geïdentificeerd zijn (Jöreskog en Sörbom, 1989). Het kan aangetoond worden dat er identificatieproblemen zijn bij de verzamelingen parameters $\{\beta_2, \sigma^2_{\eta_1}, \sigma^2_{\varepsilon_1}\}$ en $\{\sigma^2_{\varepsilon_4}, \sigma^2_{\eta_4}\}$. Hoe dat precies in zijn werk gaat, is hier niet van belang. In het geval dat de metingen op dezelfde schaal zijn uitgevoerd, is de meest natuurlijke en gangbare manier om deze onbepaaldheden te elimineren door het introduceren van de restricties $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2}$ en $\sigma^2_{\varepsilon_3} = \sigma^2_{\varepsilon_4}$. Bij de analyse van de correlatiematrix impliceert de eerste restrictie dat de betrouwbaarheden van de eerste twee toetsen gelijk zijn, de tweede restrictie impliceert dat de betrouwbaarheden van de laatste twee toetsen gelijk zijn. Het schatten van de parameters en de modeltoetsing kan plaatsvinden met behulp van standaardprogrammatuur voor lineaire structurele modellen zoals LISREL (Jöreskog & Sörbom, 1989) en EQS (Bentler, 1985). De waarde van het formuleren, schatten en toetsen van het model ligt voornamelijk in het feit van de beschikbaarheid van de programmatuur en de impliciete mogelijkheden om het model te toetsen. Daarnaast is er een zekere flexibiliteit om het model uit te breiden naar meer indicatoren voor een latente vaardigheid alsook het simultaan analyseren van verschillende latente vaardigheden.

Uiteraard zijn er naast de LISREL-benadering meer mogelijkheden om de onbekende structurele parameters te schatten. Een manier, die soelaas kan bieden in een situatie met ontbrekende waarnemingen staat beschreven in Shumway en Stoffer (1982).

10.3 Itemresponstheorie en groeiscores

In deze paragraaf werken we de bepaling van groeiscores nader uit, waarbij we een itemresponsmodel als meetmodel hanteren. Aan de hand van een concreet voorbeeld, de Schaal Vorderingen in Spellingvaardigheid (SVS) (Van den Bosch, Gillijns, Krom & Moelands, 1991), zullen we het traject voor de bepaling van groeiscores doorlopen. In tegenstelling tot bij het klassieke meetmodel, is bij itemresponsmodellen de relatie tussen de ware score of latente vaardigheid en het toetsresultaat of observaties niet lineair. Zoals zal blijken, is deze complicatie niet wezenlijk voor het bepalen van groeiscores.

10.3.1 Schaal Vorderingen in Spellingvaardigheid

Met de SVS kan men vaststellen hoe goed een leerling kan spellen in de aanvangsfase van het basisonderwijs, of anders gezegd: kan men spellingvaardigheid meten op het niveau van groep 3 en 4 van de basisschool. In deze paragraaf schetsen we summier op welke wijze dit instrument tot stand is gekomen. Bij spellen gaat het erom woorden om te zetten in schriftbeelden. Daarbij kan onderscheid gemaakt worden tussen klankzuivere en niet-klankzuivere woorden. De eerste fase van het spellingonderwijs richt zich op het correct leren schrijven van de klankzuivere woorden: je schrijft op wat je hoort. Al snel daarna komen de niet-klankzuivere woorden, de woorden waarbij er geen eenduidige relatie is tussen klank en letter, zoals bij bomen, trein, begin. Om die goed te schrijven moeten de leerlingen regels kunnen toepassen, of een woord naar analogie van een ander woord kunnen schrijven. De SVS beperkt zich tot eenvoudige klankzuivere en niet-klankzuivere woorden van een of twee lettergrepen (zie Van den Bosch e.a., 1991). De afname is klassikaal: de leerkracht leest een woord hardop voor en de leerlingen schrijven het op. De scoring is dichotoom: een correct geschreven woord levert 1 punt op en een fout geschreven woord 0 punten. In totaal bestaat het aantal opgaven van de SVS uit 173 woorden. Uit deze woorden zijn toetsen samengesteld, in totaal negen verschillende modules van elk ongeveer 20 items. In wisselende combinaties zijn deze modules op vier tijdstippen, medio en eind groep 3 en medio en eind groep 4, afgenomen bij dezelfde landelijke gestratificeerde steekproef (circa 1800 leerlingen). Het afnamedesign is al aan de orde geweest in hoofdstuk 8 en is daar weergegeven in figuur 8.5. Elke afnamegroep maakt op een tijdstip twee modules; bovendien is er voor gezorgd dat geen enkele leerling twee maal dezelfde module maakt. Dit resulteert in een design dat onvolledig is zowel op als over tijdstippen. In

equivaleerterminologie hebben we op tijdstippen met horizontaal equivaleren en over tijdstippen met verticaal equivaleren te maken. Zoals gesteld in hoofdstuk 8 komt het equivaleren neer op het calibreren van dit structurele onvolledige design met een itemresponsmodel. Bij de calibratie, dat is het schatten en toetsen van de modelparameters, is voor de SVS gebruik gemaakt van het 'One Parameter Logistic Model' (OPLM; Verhelst & Eggen, 1989). De basisvergelijking van dit model is gegeven door:

$$P(X_{vi} = x_{vi} \mid \theta_v, a_i, \beta_i) = \frac{\exp\left[a_i(\theta_v - \beta_i)x_{vi}\right]}{1 + \exp\left[a_i(\theta_v - \beta_i)\right]}.$$

In het geval van de SVS is in deze vergelijking X_{vi} een dichotome stochast bevattende de score van leerling v op item i met mogelijke waarden 0 (woord fout geschreven) en 1 (woord correct geschreven). Verder duidt θ_v de latente vaardigheid aan voor leerling v en zijn β_i en a_i respectievelijk de moeilijkheidsparameter en de discriminatie-index van item i. Voor een gedetailleerde beschrijving van dit model alsmede schattings- en modeltoets- procedures wordt verwezen naar de hoofdstukken 4 en 5. Met behulp van het OPLM bleek het mogelijk, een goede beschrijving van de SVS-data te geven. Dit resulteerde in discriminatie-indices en schattingen van de moeilijkheidsparameters voor de SVS-items. Het model werd expliciet getoetst op twee vormen van itemonzuiverheid (zie hoofdstuk 9), te weten: ethniciteit en tijdstip. Items bleken hetzelfde te functioneren voor allochtonen en autochtonen en op verschillende tijdstippen.

Nu we de items van de SVS op een schaal hebben afgebeeld, gaan we op zoek naar de nog onbekende latente vaardigheden voor de individuele leerlingen, θ_v. De itemparameters veronderstellen we in het vervolg bekend, geen onredelijke aanname gezien de omvang van de steekproef.

10.3.2 Het schatten van de latente vaardigheid

Nu de calibratie van de SVS-items met succes is afgerond, kunnen alle items in een itembank worden opgeslagen. Merk op dat er geen aanname gemaakt is over een populatieverdeling van de latente vaardigheid; de calibratie is immers uitgevoerd met CML en niet met MML (zie ook paragraaf 8.3.3). De volgende stap is het plaatsen van de individuele vaardigheden op dezelfde schaal als de items. Als vaardigheidsparameters en itemparameters op dezelfde schaal geplaatst zijn, is het meten van veranderingen in principe zonder meer mogelijk. Vaardigheden van leerlingen kunnen vergeleken worden op en over tijdstippen, en ook een terugkoppeling naar beheerste leerstof is

mogelijk door interpretatie van de itemparameters. Hoe de individuele vaardigheid geschat kan worden met een itemresponsmodel als meetmodel zullen we nu demonstreren. Wederom vergelijken we de statische en de dynamische aanpak.

Statische aanpak

Analoog aan paragraaf 10.2.2 bekijken we de tijdstippen afzonderlijk. Ook negeren we vooralsnog alle a priori kennis omtrent de populatie waartoe een leerling behoort. Op een tijdstip beschikken we voor een leerling v dus alleen over zijn toetsresultaat. In het geval dat we OPLM als meetmodel hanteren, is het toetsresultaat de som over de gemaakte items van de responsvariabele gewogen met de discriminatie-index: $s = \sum_i a_i x_{vi}$. Merk op dat het toets- resultaat s een voldoende statistiek is voor de vaardigheidsparameter θ. De vraag is nu of we de latente vaardigheid van een leerling op een tijdstip kunnen schatten uit de itemparameters en het toetsresultaat. Stel dat we de vaardigheid van een leerling opvatten als een onbekende constante, dat wil zeggen een statistische parameter die geschat moet worden. In het OPLM is het toetsresultaat een voldoende statistiek voor de vaardigheidsparameter. Een goede schatter voor de vaardigheidsparameter is de gewogen-grootste-aannemelijkheidsschatter (WML), geïntroduceerd door Warm (1989). In paragraaf 4.5 is deze schatter al besproken; hier volstaan we met het geven van de schattingsvergelijking, die wordt gegeven door het maximaliseren van de aannemelijkheidsfunctie gewogen met de toetsinformatie

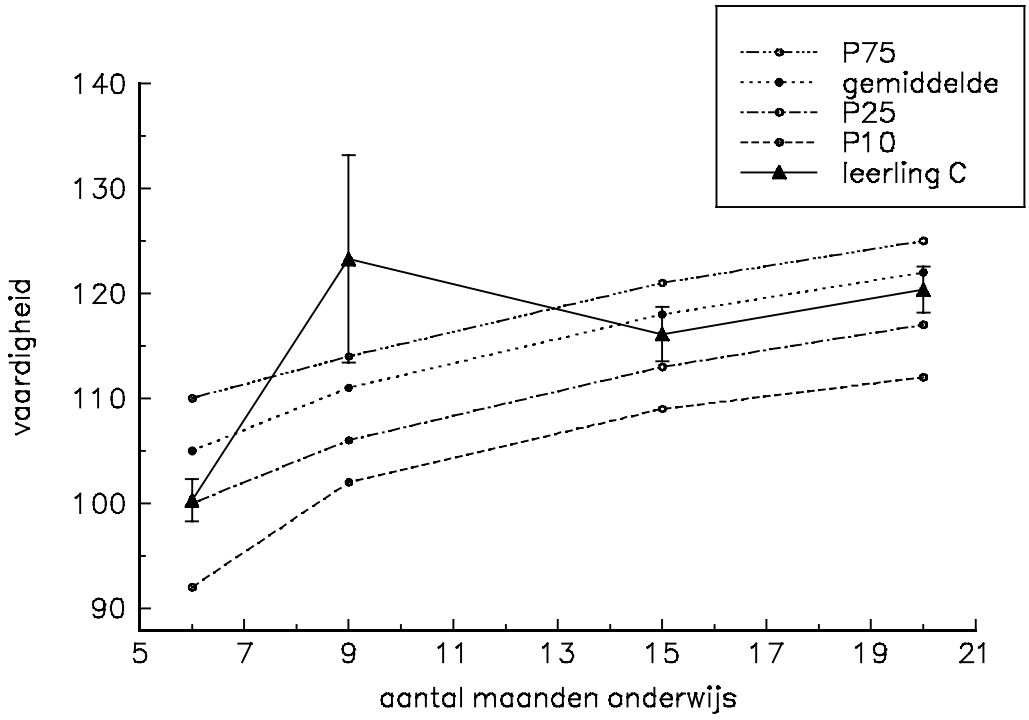
$$\max_{\theta} P(s \mid \theta) \sqrt{I(\theta)} .$$

De WML-schatter is onzes inziens de aangewezen schatter als we iemands vaardigheid opvatten als een onbekende constante. Deze schatter is immers nagenoeg zuiver op het individuele niveau en bestaat ook voor leerlingen die alles fout dan wel goed hebben, dit in tegenstelling tot de gewone grootste-aannemelijkheidsschatter. De WML-schatter voor de latente vaardigheid in een itemresponsmodel is het equivalent van de geobserveerde-score-schatter van de ware score in het klassieke meetmodel. In tegenstelling tot de geobserveerde-score-schatter uit het klassieke meetmodel is de WML-schatter een niet-lineaire transformatie van het toetsresultaat s. Uiteraard hoort bij de WML-schatter een schattingsfoutvariantie. De schattingsfoutvariantie van de geobserveerde-score-schatter in het klassieke meetmodel is gelijk aan de meetfoutvariantie en onafhankelijk van de ware score van een leerling, en is voor elke geobserveerde score even groot. Daarentegen is de schattingsfoutvariantie van de WML- schatter

afhankelijk van de latente vaardigheid en dus voor leerlingen met een ongelijk toetsresultaat verschillend.

Vanwege de eigenschap van zuiverheid van de WML-schatter is het mogelijk, populatie- karakteristieken te achterhalen als percentielen en gemiddelden. Deze populatiekarakteristieken kunnen dan vervolgens dienen als referentiegegevens voor individuele resultaten. Stel dat we voor de SVS referentiegegevens zoals gemiddelden en percentielen willen bepalen voor de Nederlandse populatie leerlingen per tijdstip, dan kan dat simpel door bijvoorbeeld de WML-schattingen in de steekproef te middelen, of bij het bepalen van percentielen de WML-schattingen in de steekproef te sorteren naar oplopende grootte en die waarden te kiezen die corresponderen met de percentages. Daar de steekproef in het voorbeeld van de SVS gestratificeerd was naar schoolgewicht (zie ook paragraaf 7.1), diende uiteraard een weging plaats te vinden naar de Nederlandse populatie. In figuur 10.9 zijn voor de Nederlandse populatie leerlingen per tijdstip het gemiddelde en de percentielen 10, 25 en 75 weergegeven. Tevens zijn in figuur 10.9 voor leerling C de WML-schatting op de vier momenten weergegeven.

Met behulp van de referentiegegevens kunnen we nu bepalen hoe goed een leerling het doet ten opzichte van de groep op de vier meetmomenten. Kijken we naar de WML-schattingen van leerling C, dan kunnen we constateren dat na zes maanden onderwijs de vaardigheid van deze leerling rond percentiel 25 ligt, na negen maanden onderwijs ver boven percentiel 75 en terugvalt onder het gemiddelde na vijftien en twintig maanden onderwijs. Rond de schat-tingen voor leerling C is een betrouwbaar-heidsinterval aangegeven, plus en min een standaardafwijking van de schattingsfout, de verticale lijntjes in figuur 10.9. De orde van grootte van de betrouwbaarheidsintervallen is ongeveer 5 punten op de schaal voor de SVS, met uitzondering voor tijdstip 2, dat is na 9 maanden onderwijs; daar omvat het interval circa



Figuur 10.9 Gemiddelden en percentielen (P10, P25 en P75) voor de Nederlandse populatie in groep 3 en 4 van de basisschool voor de SVS en de WML-schattingen voor leerling C

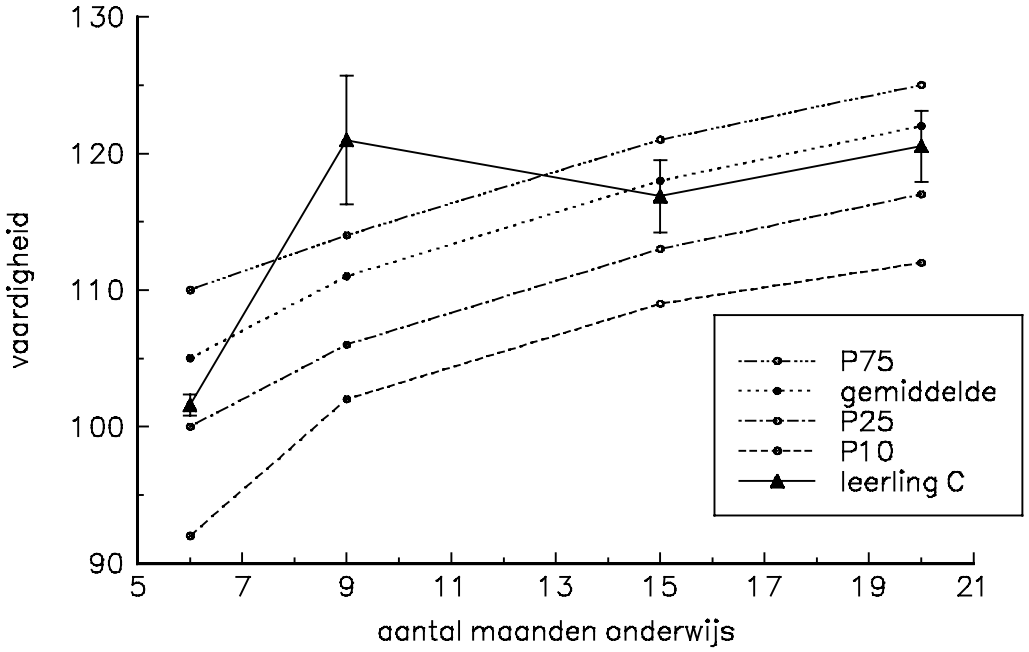
20 punten. Op tijdstip 2 hebben we de vaardigheid van leerling C dus zeer onnauwkeurig gemeten. Dit is problematisch als men resultaten wil interpreteren of conclusies verbinden aan de ontwikkeling van leerling C met betrekking tot spellingvaardigheid. In de praktijk van het onderwijs is het beeld als geschetst voor leerling C, eerder regel dan uitzondering. Deze fluctuaties van de vaardigheid in de tijd voor een leerling is voor het leeuwedeel te wijten aan de vaak zeer onbetrouwbare metingen.

In het kader van de itemresponstheorie zijn er diverse mogelijkheden om de nauwkeurigheid van de metingen te vergroten. Te denken valt aan vormen van adaptief toetsen. We komen hier straks op terug. Een andere mogelijkheid is de schatting van de latente vaardigheid van een leerling niet alleen te laten afhangen van zijn eigen toetsresultaat, maar ook van informatie over de groep waartoe deze leerling behoort. Merk de analogie met de Kelley schatter in paragraaf 10.2.2 op. Het equivalent van de Kelley-schatter uit het klassieke meetmodel in de itemresponstheorie is de 'expected a posteriori' of EAP-schatter. De EAP-schatter is al besproken in hoofdstuk 4; hier volstaan we alleen met de schattingsvergelijking:

$$\mathscr{E}(\theta \mid s) = \frac{\int \theta P(s \mid \theta)\, g(\theta)\, d\theta}{\int P(s \mid \theta)\, g(\theta)\, d(\theta)}, \qquad (10.20)$$

waarbij $g(\theta)$, de kansdichtheidsfunctie van θ is in de populatie, dus de populatie-informatie met betrekking tot θ.

Om de EAP-schatter uit te kunnen rekenen moeten we over populatie-informatie $g(\theta)$ beschikken. Daartoe zullen we $g(\theta)$ moeten specificeren. Gebruikelijk is, hiervoor de normale verdeling te kiezen. Gemiddelde en variantie van deze a priori verdeling zullen we moeten schatten. Schattingen kunnen we onder andere verkrijgen met behulp van de MML- methode, besproken in hoofdstuk 4, of door statistiek te bedrijven met de WML-schattingen (Verhelst & Kamphuis, 1989; Hoijtink & Boomsma, 1991). Hier volstaan we met het geven van schattingen van deze verdelingen op de vier tijdstippen. Deze zijn voor het gemiddelde respectievelijk 105.2, 111.3, 117.3 en 121 en voor de varianties respectievelijk 101.6, 53.6, 51.1 en 56.7. In wezen zijn dit de a priori schattingen uit paragraaf 10.2.2, waarbij men het



Figuur 10.10
EAP-schattingen voor leerling C

gemiddelde kan opvatten als schatter en de variantie als schattingsfoutvariantie. In figuur 10.10 zijn voor leerling C de EAP-schattingen en de betrouwbaarheidsintervallen (plus en min één standaardafwijkingen van de schattingsfout) weergegeven.

Men kan constateren dat op alle tijdstippen de WML-schattingen in de richting van het populatiegemiddelde zijn opgeschoven. De verschuiving is het grootst op tijdstip 2 waar de WML-schatting het meest onbetrouwbaar was. Ook kan geconstateerd worden dat

in dit geval de schattingsfout bij de EAP-schattingen kleiner is dan bij de WML-schattingen. Dit hoeft niet altijd het geval te zijn.

Resumerend kunnen we stellen dat bij de statische benadering van groeiscores de schatters uit de klassieke testtheorie equivalenten hebben in de itemresponstheorie.

Dynamische benadering

Ook de drie besproken schatters bij de dynamische benadering in paragraaf 10.2.3, de predictieve, Kalmanfilter- en gladgestreken Kalmanfilterschatters, hebben hun equivalenten in de itemresponstheorie. Merk op dat met betrekking tot het groeimodel, op populatieniveau geformuleerd, er niets verandert als we in plaats van de klassieke testtheorie de itemresponstheorie als meetmodel hanteren. Het groeimodel beschrijft immers niets anders dan de ontwikkeling van de latente vaardigheid in de tijd ongeacht de wijze waarop we die vaardigheid ook trachten te meten. Dit houdt in dat de predictieve schatter voor beide modellen dezelfde vorm heeft, alleen de schatting die we invullen in bijvoorbeeld (10.14) is anders en wordt nu bepaald door het gebruikte meetmodel. Uitgaande van hetzelfde autoregressieve groeimodel als besproken in paragraaf 10.2.3, kan de procedure voor het verkrijgen van de dynamische schatters in de volgende stappen uiteengelegd worden:

(1) Bepaal op het eerste tijdstip $\mathscr{E}(\theta_1 \mid s_1, \mu_{\theta_1}, \sigma^2_{\theta_1})$, dat is de EAP-schatter gegeven het toetsresultaat s_1 en de marginale verdeling van θ op tijdstip 1 met gemiddelde μ_{θ_1} en variantie $\sigma^2_{\theta_1}$, en bijbehorende schattingsfoutvariantie (Kalmanfilterschatter).

(2) Deze conditionele verwachting en schattingsfoutvariantie substitueren we in de pre- dictievergelijking 10.14. Nu beschikken we over de predictieve schatter en schattingsfoutvariantie op meetmoment 2.

(3) Bepaal de Kalmanfilterschatting op tijdstip 2, dat is de EAP-schatter gegeven toetsresultaat, s_2, en de predictieve schatter en schattingsfoutvariantie uit stap 2.

(4) Herhaal stap 2 en 3 tot alle meetmomenten verwerkt zijn.

(5) Bepaal met behulp van de nu beschikbare Kalmanfilterschattingen en schattings-foutvarianties de gladgestreken schattingen en bijbehorende schattingsfoutvarian-ties.

In de klassieke testtheorie kwam de combinatie van populatieinformatie en toetsresultaat in essentie neer op het combineren van twee onafhankelijke schatters, de geobserveerde-score-schatter en de predictieve schatter tot de Kelley-schatter. In de itemresponstheorie vervult de EAP-schatter de rol van de Kelley-schatter.

De vraag resteert hoe we de gemiddelden en de covariantiematrix van de latente vaardigheid op populatieniveau kunnen schatten. Het voert te ver hier op in te gaan; we volstaan met een verwijzing naar Kamphuis en Engelen (in voorbereiding). In het voorbeeld van de SVS is een autoregressief model van de eerste orde geschat voor de vier meetmomenten:

$$\theta_t = a_t + b_t \theta_{t-1} + \zeta_t \qquad t = 2, 3, 4,$$

waarbij t de tijdstipindex, a en b de regressiecoëfficiënten en ζ_t een storingsvariabele met verwachting 0 en variantie Ψ_t (onverklaarde variantie op een tijdstip t) is. Schattingen voor de parameters in deze vergelijkingen staan vermeld in tabel 10.4.

Gemiddeld groeit de populatie circa 6 punten tussen tijdstippen, uitgezonderd voor het laatste tijdstip. De voorspellingen van tijdstip naar tijdstip verklaren respectievelijk 62%, 70% en 81% van de variantie op de desbetreffende tijdstippen. Laten we eens zien wat de consequenties zijn als we dit groeimodel toepassen op leerling C. In figuur 10.11 zijn de Kalmanfilterschattingen voor leerling C weergegeven en in figuur 10.12 de gladgestreken Kalmanfilterschattingen. Als we kijken naar tijdstip 2, dan kunnen we constateren dat de Kalmanfilterschatter nog meer dan de EAP-schatter de schaalscore heeft verminderd, respectievelijk 116.31 en 120.98.

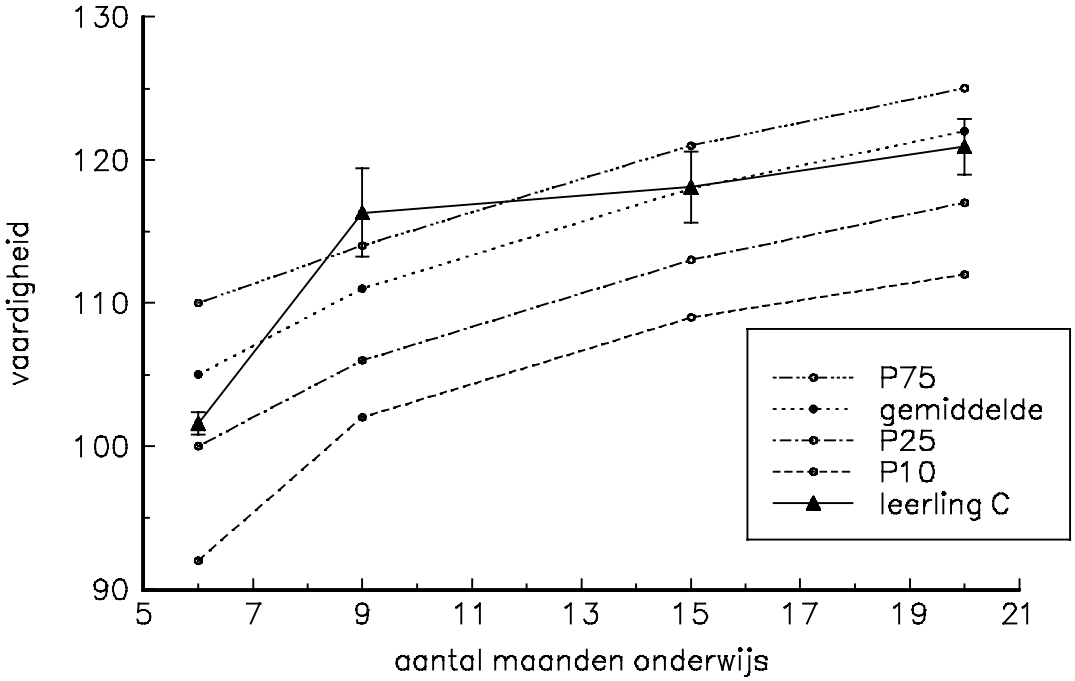
Tabel 10.4

Schattingen van de parameters van het SVS groeimodel met tussen haakjes het aantal maanden onderwijs

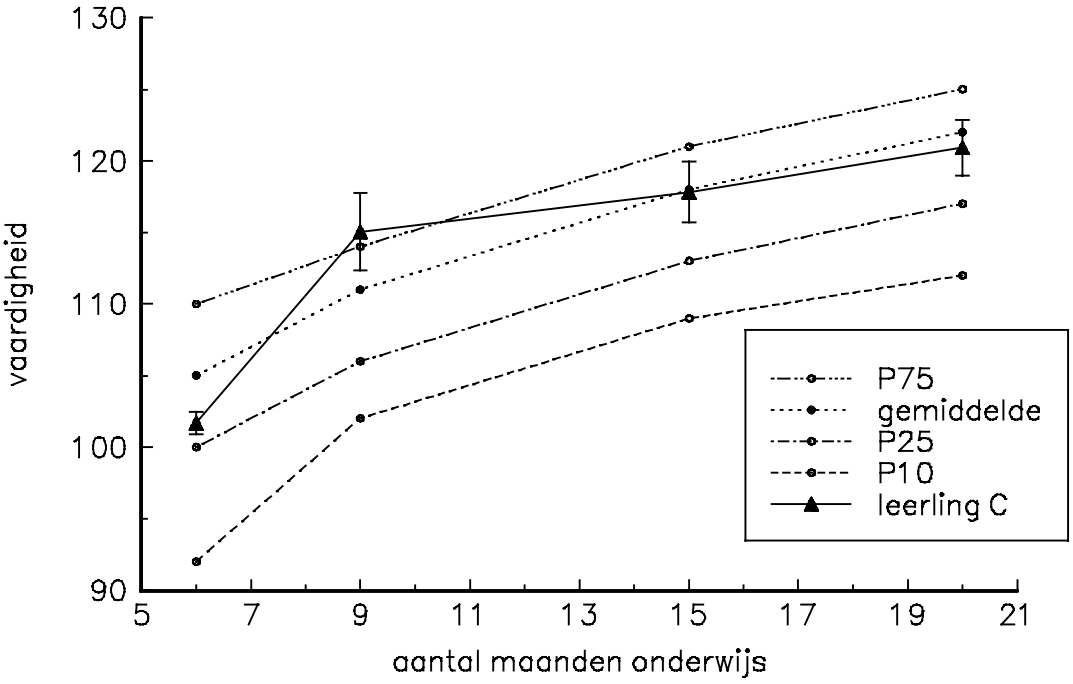
	tijdstip			
parame-ter	1(6)	2(9)	3(15)	4(20)
μ_θ	105.15	111.32	117.34	120.95
σ^2_θ	101.60	53.58	51.10	56.74
Ψ		20.18	15.52	10.53
a		51.02	26.62	9.38
b		.57	.81	.95

De predictieve schatting en schattingsfout, de a priori kennis op tijdstip 2, bedroeg 109.27 en 4.51 (niet weergegeven in figuur 10.11). Bij de EAP-schatter daarentegen was de a priori kennis gebaseerd op een gemiddelde 111.32 en een standaarddeviatie van 7.32. Ook constateren we weer dat toevoegen van informatie uit het groeimodel de schattingsfouten reduceert. De gladgestreken schatting op tijdstip 2 voor leerling C ligt in vergelijking met de Kalmanfilterschatting meer in lijn met de andere schattingen.

Ook constateren we weer dat de standaardschattingsfouten van de gladgestreken Kalmanfilterschattingen iets kleiner uitvallen dan die van de Kalmanfilterschattingen.



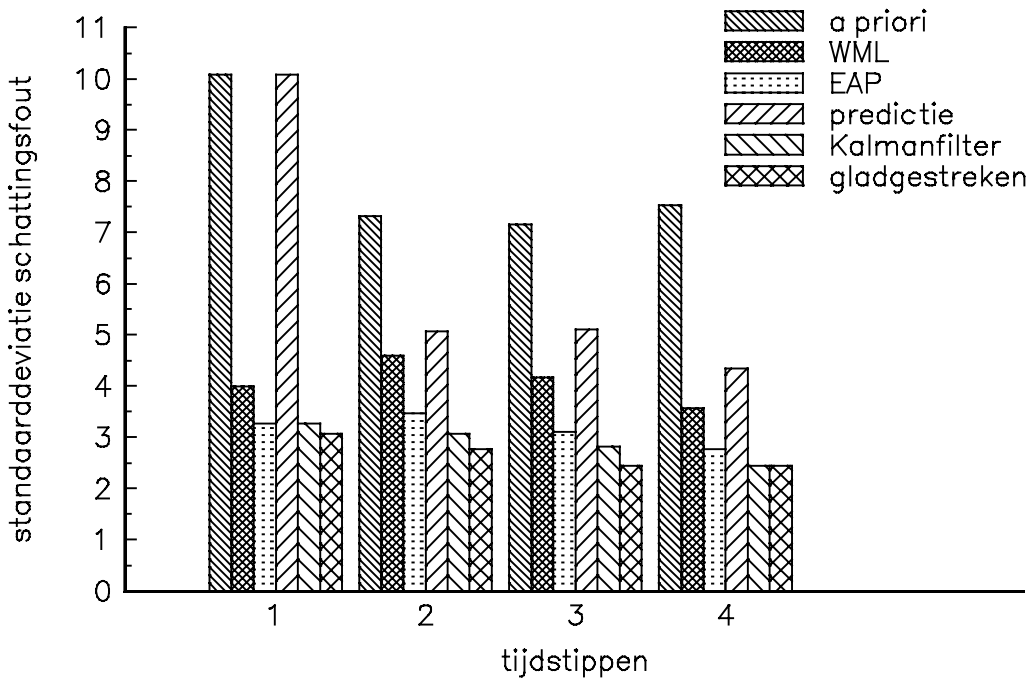
Figuur 10.11
Kalmanfilterschattingen voor leerling C



Figuur 10.12 Gladgestreken Kalmanfilterschattingen voor leerling C

Evaluatie statische en dynamische benadering

De conclusies die getrokken zijn in de evaluatie van de statische en dynamische benadering bij het bepalen van individuele vaardigheden in paragraaf 10.2.4 gelden natuurlijk onverkort in de situatie waarin een itemresponsmodel wordt gebruikt als meetmodel. In het voorbeeld van de SVS beschikken we echter niet over de ware vaardigheden van de individuele leerling zoals in het voorbeeld van knikkervaardigheid. Dus, om de in deze paragraaf besproken statische en dynamische schatters te evalueren kunnen we alleen terugvallen op de statistische eigenschappen van deze schatters. Daar alle besproken schatters wederom zuiver zijn in de populatie, beperken we ons ook deze keer tot een vergelijking van een maat voor de spreiding van de schattingsfout van de diverse schatters. In figuur 10.13 is de gemiddelde standaardafwijking van de schattingsfout voor de diverse schatters op de verschillende tijdstippen weergegeven. We vergelijken eerst de standaardafwijkingen van de schattingsfout van de cross-sectionele schatters. De EAP-schatter heeft op alle tijdstippen de kleinste standaardafwijking, gevolgd door de WML-schatter en de a priori schatter. Verder valt op dat de stan-



Figuur 10.13 Gemiddelde standaarddeviatie van de schattingsfout voor de diverse schatters op de vier tijdstippen voor de leerlingen in de SVS-steekproef (n = 1800)

daardafwijking van de a priori schatter op het eerste tijdstip groter is dan op de volgende tijdstippen (circa 10 versus circa 7). Aanvankelijke verschillen in spellingvaardigheid in de populatie worden blijkbaar door het effect van het onderwijs deels geneutraliseerd. Ook constateren we dat de standaardafwijking van de WML-schatter op het tweede tijdstip in vergelijking met de andere tijdstippen het grootst is. De toetsmodules die zijn afgenomen op het tweede tijdstip leverden de minste informatie op over de spellingvaardigheid. Met andere woorden: deze modules zijn niet op maat gesneden voor de populatie op dat tijdstip. Bezien we de dynamische schatters, dan is het beeld niet anders dan beschreven in paragraaf 10.2.4: de gladgestreken Kalmanfilterschattingen zijn het meest nauwkeurig, gevolgd door de Kalmanfilterschattingen en op afstand de predictieve schattingen.

Ook hier constateren we dat de dynamische schatters hun statische equivalenten overtreffen als het gaat om de meetnauwkeurigheid. De mate waarin, wordt bepaald door de precisie van de meetresultaten en de mate van nauwkeurigheid van de predicties.

10.4 Epiloog

In dit hoofdstuk is het meten van veranderingen en het bepalen groeiscores behandeld. De kern van het verhaal ligt besloten in de vraag: Hoe combineren we informatie uit twee bronnen, groei- en meetmodel, tot één vaardigheidsschatting? We zagen dat het mogelijk was om met behulp van een groeimodel iemands vaardigheid te voorspellen op een bepaald tijdstip. Bovendien konden we op dat tijdstip de actuele meting met behulp van een meetmodel omzetten in een schatting van de vaardigheid. Groei- en meetmodel leverden dus beiden een indicatie op over iemands vaardigheid, welke gecombineerd konden worden tot één schatting. Afhankelijk van het gekozen meet- en/of groeimodel en de keuze hoe men de vaardigheid beziet, als een onbekende parameter of als een toevalsvariabele, ziet de schatter er anders uit. Welke schatter men prefereert, is vaak een persoonlijke keuze. De meest informatieve schatter is de gladgestreken Kalmanfilterschatter. De minst informatieve schatter is in de klassieke testtheorie de geobserveerde score en in de itemresponstheorie de WML-schatter. De keuze voor de minst informatie schatter wordt vaak gemotiveerd door te stellen dat men geen a priori informatie wil meenemen in de schatting van de vaardigheid omwille van de eerlijkheid. Met andere woorden, de schatting van de vaardigheid mag alleen berusten op het meetresultaat en niet mede bepaald worden door eerdere meetresultaten of door de groep waartoe iemand behoort. Dit lijkt een nobel standpunt. Statistisch bezien is dit standpunt echter onrealistisch daar alle ingrediënten van deze schatters populatie afhankelijk zijn. In de klassieke testtheorie zijn de indexen voor de betrouwbaarheid zonder de definitie van een populatie betekenisloos. In de itemresponstheorie hebben de itemparameters altijd betrekking op een populatie, ook al bestaan er fraaie schattingsprocedures voor de itemparameters die steekproefonafhankelijk zijn. In de onderwijspraktijk levert dit standpunt dan ook problemen op: Hoe moeten we onbetrouwbare schattingen van de vaardigheid voor een leerling, die excessief fluctueren in de tijd, interpreteren? Dit excessief fluctueren van de vaardigheid in de tijd op het individuele niveau, door Rubin (1980) in een ander kader het "bouncing beta problem" genoemd, kan onderdrukt worden door populatie-informatie (groeimodel) te gebruiken bij de schattingen van iemands vaardigheid. Tevens reduceert dit deels de onbetrouwbaarheid van de schattingen. Een andere mogelijkheid om de onbetrouwbaarheid van de schattingen te reduceren, kan gevonden worden in de toepassing van betere meetprocedures. Met als uitgangspunt een schatter die informatie uit groei- en meetmodel combineert, bezien we welke mogelijkheden er zijn om de nauwkeurigheid van de schattingen te verhogen.

Eerst kijken we op het niveau van de populatie naar het groeimodel. In de voorbeelden die gebruikt zijn in dit hoofdstuk werd groei voor één vaardigheid gemoduleerd middels een simpel autoregressief model van de eerste orde waarbij één populatie werd

verondersteld. In de praktijk zal een dergelijke aanname waarschijnlijk een te grove benadering van de werkelijkheid zijn. Realistischer is het te veronderstellen dat er subpopulaties of groepen zijn te onderscheiden waarbij de groei verschillend verloopt. Denkbaar is ook dat we niet kunnen volstaan met een eerste orde autoregressief groeimodel, maar dat er andere modellen te vinden zijn die een betere beschrijving van de data opleveren. In de praktijk zullen we dus moeten onderzoeken, welk groeimodel we kiezen voor wie. Naast modelselectie dienen de modellen uiteraard naar behoren getoetst te worden. Om groepen op te sporen waarvoor groei verschillend verloopt zijn er een aantal procedures denkbaar. Een eerste procedure zou kunnen starten met een opdeling van de populatie naar achtergrondkenmerken. Men zou bijvoorbeeld na kunnen gaan of groei anders gemodelleerd dient te worden voor meisjes en jongens. Een andere mogelijkheid zou kunnen zijn een latente klasse analyse uit te voeren. Bij deze benadering vormen personen die hetzelfde groeipatroon hebben één (latente) klasse. De problemen bij deze laatste benadering zijn echter legio; vooralsnog is deze benadering dan ook toekomstmuziek.

De crux van het modelleren van groei is de voorspellingen zo nauwkeurig mogelijk te krijgen. Daarom is ook additionele informatie, bijvoorbeeld informatie met betrekking tot andere vaardigheden, bruikbaar om de predicties te verbeteren. Oud en Mommers (1988) gebruiken een longitudinaal verklaringsmodel voor de samenhang tussen de vaardigheden technisch lezen, begrijpend lezen en spellen. In dit model kan bij de predictie van spellingvaardigheid op een zeker tijdstip, informatie van de vaardigheden technisch lezen en begrijpend lezen worden verbeterd.

De mogelijkheden om de onbetrouwbaarheid van de schattingen van de vaardigheid te reduceren met behulp van het meetmodel zijn sterk afhankelijk van het gebruikte meetmodel. Merk ook op dat reductie van de schattingsfouten alleen kan plaatsvinden bij een nieuwe afname, reeds afgenomen toetsen kunnen niet meer bijgesteld worden. Laten we eens aannemen dat er aan de hand van een longitudinale gegevensverzameling een groeimodel voor een bepaalde populatie geschat hebben. Het is nu in principe mogelijk de meetprocedure voor toekomstige afnames te verfijnen op basis van de reeds beschikbare gegevens. Wel moeten we dan bedenken dat we bepaalde assumpties moeten maken, bijvoorbeeld dat de leerlingen bij de toekomstige afname beschouwd kunnen worden een steekproef uit oorspronkelijke populatie of dat de itemparameters in een itemresponsmodel constant blijven in de tijd. Zeker in een longitudinale context, waarbij de tijdspannes vaak groot zijn, is het wenselijk deze assumpties te controleren. Het is bijvoorbeeld denkbaar dat itemparameters als gevolg van onderwijskundige ontwikkelingen, door de loop der jaren veranderen. Stel dat er voor een leerling een vaardigheidsschatting beschikbaar is op een bepaald tijdstip. Met behulp van het

groeimodel is het mogelijk te voorspellen hoe vaardig de leerling op een volgend tijdstip zal zijn. Gegeven deze voorspelling, kunnen we dan voor deze leerling een toets op 'maat' kiezen, dat wil zeggen een toets kiezen die de meetfout minimaliseert. Hoe we toetsen op maat kunnen samenstellen wordt besproken in hoofdstuk 11. Ook kunnen predicties van de vaardigheid gebruikt worden als startwaarden in adaptieve toetsprocedures, dat is biedt opgaven aan met een moeilijkheid in de buurt van de lopende schatting van de vaardigheid. Merk op dat de mogelijkheden van toetsen op maat sterk bepaald zijn door het gebruikte meetmodel. Al met al bieden itemresponsmodellen in zijn algemeenheid meer mogelijkheden voor verfijnde toetsprocedures dan het klassieke meetmodel.

Het belang van de keuze van een geschikt meet- en groeimodel bij het meten van veranderingen kan niet genoeg benadrukt worden. Zowel het meetmodel als het groeimodel kunnen in belangrijke mate bijdragen aan de reductie van de onbetrouwbaarheid van de vaardigheidsschattingen voor individuele leerlingen. Als we de vaardigheid van de leerlingen in de tijd nauwkeurig kunnen bepalen, kunnen we ook het probleem van een verfijnd referentiekader (zie paragraaf 10.1.3) aanpakken. We kunnen dan de individuele groei nauwkeurig afzetten tegen relevante andere individuen, groepen en populaties maar ook tegen onderwijsinhoudelijke criteria. Maar dan moet het ook mogelijk zijn om ongewenste ontwikkelingen of problemen te signaleren, bijvoorbeeld achterstand. Tenslotte nog een laatste opmerking. De signalering van problemen alleen is niet voldoende; diagnostisering van problemen en de ontwikkeling van hulpprogramma's voor achterstanden verdienen de nodige zorg en aandacht. Hopelijk biedt het hier geschetste kader voor het meten van veranderingen, waarbij meet- en groeimodel gekoppeld zijn, voldoende aanknopingspunten voor de gerichte ontwikkeling van diagnose- en hulpmateriaal.